

4/5/1 (Item 1 from file: 351)  
DIALOG(R) File 351: Derwent WPI  
(c) 2004 Thomson Derwent. All rts. reserv.

012386060 \*\*Image available\*\*  
WPI Acc No: 1999-192167/199917  
XRPX Acc No: N99-140742

**Large capacity multi-class core ATM switch architecture**

Patent Assignee: NEC CORP (NIDE ); NEC USA INC (NIDE )

Inventor: FAN R; MARK B; RAMAMURTHY G; MARK B L

Number of Countries: 027 Number of Patents: 003

**Patent Family:**

Patent No	Kind	Date	Applicat No	Kind	Date	Week
EP 901302	A2	19990310	EP 98107930	A	19980430	199917 B
JP 11088374	A	19990330	JP 9892088	A	19980403	199923
US 6324165	B1	20011127	US 97923978	A	19970905	200175

Priority Applications (No Type Date): US 97923978 A 19970905

**Patent Details:**

Patent No	Kind	Lan Pg	Main IPC	Filing Notes
-----------	------	--------	----------	--------------

EP 901302	A2	E 43	H04Q-011/04	
-----------	----	------	-------------	--

Designated States (Regional): AL AT BE CH CY DE DK ES FI FR GB GR IE IT

LI LT LU LV MC MK NL PT RO SE SI

JP 11088374	A	29	H04L-012/28
-------------	---	----	-------------

US 6324165	B1		H04L-012/26
------------	----	--	-------------

Abstract (Basic): EP 901302 A2

NOVELTY - The large capacity triple-buffered switch is preferably a single stage switch which may be used as a switching element in a still larger capacity multi-stage switch. The core switch module (10) consists of 16 input ports (IP), 16 output ports (OP) and one multicast output port (MOP), all connected to a high speed terminal bus (20). The core module interconnects the input/output modules to large buffers and intelligent/management mechanisms.

DETAILED DESCRIPTION - INDEPENDENT CLAIMS are included for an ATM switch scheduler and a buffer

USE - For servicing requests originated from various classes of sources in ATM network.

ADVANTAGE - Can flexibly vary bandwidth of switch and cope with multiple traffic class requirements

DESCRIPTION OF DRAWING(S) - The drawing is a diagram depicting architecture of core switch module according to preferred embodiment of present invention.

Core switch module (10)

Input ports (IP)

Output ports (OP)

Multicast output port (MOP)

High speed terminal bus (20)

pp; 43 DwgNo 1/18

Title Terms: CAPACITY; MULTI; CLASS; CORE; ATM; SWITCH; ARCHITECTURE

Derwent Class: W01

International Patent Class (Main): H04L-012/26; H04L-012/28; H04Q-011/04

International Patent Class (Additional): H04L-012/56; H04Q-003/00

File Segment: EPI

4/5/2 (Item 1 from file: 347)  
DIALOG(R) File 347: JAPIO  
(c) 2004 JPO & JAPIO. All rts. reserv.

06146834 \*\*Image available\*\*

**LARGE CAPACITY MULTI-CLASS CORE ATM SWITCH ARCHITECTURE**

PUB. NO.: 11-088374 A]

PUBLISHED: March 30, 1999 (19990330)

INVENTOR(s): FAN RUIXUE

MARK BRIAN  
RAMAMURTHY GOPALAKRISHNAN  
APPLICANT(s): NEC CORP  
APPL. NO.: 10-092088 [JP 9892088]  
FILED: April 03, 1998 (19980403)  
PRIORITY: 923978 [US 923978], US (United States of America), September  
05, 1997 (19970905)  
INTL CLASS: H04L-012/28; H04Q-003/00

#### ABSTRACT

PROBLEM TO BE SOLVED: To provide a large capacity ATM core switch structure for supporting plural traffic classes and service quality guarantee.

SOLUTION: A switch efficiently adjusts real-time and non-real-time multi-cast flows and the switch is constituted of a high-speed core module 10 for mutually connecting input/output modules provided with a large capacity buffer and an intelligent scheduling/buffer management mechanism. Scheduling is realized by using new dynamic rate control for controlling internal congestion and achieving fair throughput performance between the flows competing at a switch bottleneck part. In a dynamic rate control system, the flow is rate-controlled corresponding to congestion information monitored at the bottleneck part inside the switch. For respective switch flows, a rate for which a dynamic rate component for fairly dividing an unused band width is added to a minimum service rate is guaranteed.

COPYRIGHT: (C)1999,JPO

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-88374

(43) 公開日 平成11年(1999) 3月30日

(51) Int.Cl.<sup>6</sup>

識別記号

F I

H 0 4 L 12/28

H 0 4 L 11/20

G

H 0 4 Q 3/00

H 0 4 Q 3/00

H 0 4 L 11/20

H

審査請求 有 請求項の数27 O L (全 29 頁)

(21) 出願番号 特願平10-92088

(22) 出願日 平成10年(1998) 4月3日

(31) 優先権主張番号 08/923978

(32) 優先日 1997年9月5日

(33) 優先権主張国 米国 (US)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 ルイクシュー ファン

アメリカ合衆国, ニュージャージー

08540, プリンストン, 4 インディペン

デンス ウエイ, エヌ・イー・シー・ユ

ー・エス・エー・インク内

(74) 代理人 弁理士 後藤 洋介 (外1名)

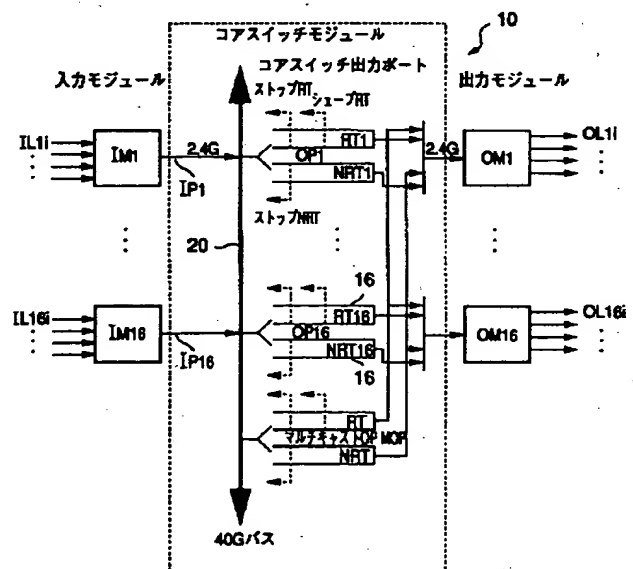
最終頁に続く

(54) 【発明の名称】 大容量マルチクラスコアATMスイッチアーキテクチャ

(57) 【要約】 (修正有)

【課題】 複数のトラフィッククラス及びサービス品質保証をサポートする大容量ATMコアスイッチ構造を提供する。

【解決手段】 スイッチは、リアルタイム及びノンリアルタイムマルチキャストフローを効率的に調整する。スイッチは、大容量バッファを備えた入力／出力モジュールを相互接続する高速コアモジュールと、インテリジェントスケジューリング／バッファ管理メカニズムとで構成される。スケジューリングは、内部輻輳を制御し、スイッチボトルネック部で競合するフロー間に公平なスループット性能を達成する新規な動的レート制御を用いて実現できる。動的レート制御方式では、フローは、スイッチ内のボトルネック部で監視される輻輳情報に応じてレート制御される。各スイッチフローには、最低サービスレートに、未使用帯域幅を公平に配分する動的レート成分を加えたレートが、保証される。



## 【特許請求の範囲】

【請求項1】 コアスイッチと、上記コアスイッチの入力側に接続された複数個の入力モジュールと、上記コアスイッチの出力側に接続された複数個の出力モジュールとを備え、

上記コアスイッチは、  
TDMバスと、

上記TDMバスに接続された複数個の入力ポートと、  
上記TDMバスに接続された複数個の出力バッファと、  
上記出力バッファにそれぞれ接続された複数個の出力ポ  
ートと、

上記出力ポートの各々に接続されたマルチキャスト出力  
バッファとを備え、

上記各入力モジュールは、

上記出力ポートの数に対応して設けられ、且つ、それぞ  
れ複数個の入力バッファを有する出力ポートプレーン  
と、

上記入力バッファにおいてセルをスケジューリングする  
ための入力モジュールスケジューラとを備え、

上記出力モジュールの各々は、

複数個の出力回線プレーンであって、各々が、出力回線  
に連結された複数個の出力回線バッファを有する出力回  
線プレーンと、

上記出力バッファにおいてセルをスケジューリングする  
ための出力モジュールスケジューラとを備えた、異なる  
サービス品質要求を有する異なるクラスのストリームを  
サポートできるATMスイッチ。

【請求項2】 請求項1において、上記コアスイッチの  
上記出力バッファの各々は、各出力バッファ内のセルの  
レベルが第1のしきい値に到達するとシェーピング信号  
を発生し、各出力バッファ内のセルのレベルが第2のし  
きい値に到達するとストップ信号を発生するオーバーフ  
ロー制御部を備えたことを特徴とするATMスイッチ。

【請求項3】 請求項1において、上記複数個の入力バ  
ッファ内の各セルに対してキュータイムスタンプを付与  
するコネクション受付制御部と、

現在時刻を与えるタイマとをさらに備え、

上記入力モジュールスケジューラは、さらに、上記キュー  
タイムスタンプと上記現在時刻を比較し、現在時刻に  
等しいキュータイムスタンプを有する上記セルの各々を  
スケジューリングするコンパレータを含むことを特徴と  
するATMスイッチ。

【請求項4】 請求項1において、上記コアスイッチの  
上記複数個の出力バッファとして、リアルタイム出力バ  
ッファが設けられると共に、上記コアスイッチは、更  
に、

上記リアルタイム出力バッファの数に対応した数だけ設  
けられ、上記出力ポートの一つにそれぞれ接続された複  
数個のノンリアルタイム出力バッファと、

上記出力ポートの各々に接続されたノンリアルタイムマ

ルチキャストバッファとを備えていることを特徴とする  
ATMスイッチ。

【請求項5】 請求項3において、上記入力モジュール  
スケジューラ及び上記出力モジュールスケジューラの各  
々は、

上記キュータイムスタンプを記憶し、上記コンパレータ  
に上記キュータイムスタンプを提供するタイムスタンプ  
記憶装置と、

サービス資格のあるセルの伝送レートをシェーピングす  
るための、上記コンパレータに接続された仮想レートシ  
ューピング装置と、

サービス資格のあるセルの数をカウントするための、複  
数個の仮想キューカウンタと、

サービス資格のあるセルをスケジューリングするための  
サービススケジューリング装置と、

上記タイムスタンプを動的に更新するための計算エンジ  
ンとをさらに備えたことを特徴とするATMスイッチ。

【請求項6】 請求項5において、上記出力モジュール  
から上記入力モジュールへのレートフィードバックをさ  
らに備えたことを特徴とするATMスイッチ。

【請求項7】 コアスイッチと、上記コアスイッチの入  
力側に接続された複数個の入力モジュールと、上記コア  
スイッチの出力側に接続された複数個の出力モジュール  
とを備え、

上記コアスイッチは、

TDMバスと、

上記TDMバスに接続された複数個の入力ポートと、

上記TDMバスに接続された複数個の出力バッファと、

上記出力バッファにそれぞれ接続された複数個の出力ポ  
ートと、

上記出力ポートの各々に接続されたマルチキャスト出力  
バッファとを備え、

上記出力モジュールの各々は、複数個の出力回線を有  
し、且つ、

複数個の出力回線プレーンであって、各々が、上記出力  
回線に連結された複数個の出力回線バッファを有する出  
力回線プレーンと、

上記出力バッファにおいてセルをスケジューリングする  
ための出力モジュールスケジューラとを備え、

上記各入力モジュールは、

上記出力ポートの数に対応して設けられた複数個の出力  
ポートプレーンであって、それぞれ、上記出力回線の数  
に対応する複数個の出力回線プレーンを有し、上記出力  
回線プレーンの各々が複数個の入力バッファを有する出  
力ポートプレーンと、

上記入力バッファにおいてセルをスケジューリングする  
ための入力モジュールスケジューラとを備えた、異なる  
サービス品質要求を有する異なるクラスのストリームを  
サポートできるATMスイッチ。

【請求項8】 請求項7において、上記コアスイッチの

## 3

上記出力バッファの各々は、各出力バッファ内のセルのレベルが第1のしきい値に到達するとシェーピング信号を発生し、各出力バッファ内のセルのレベルが第2のしきい値に到達するとストップ信号を発生するオーバーフロー制御部を備えたことを特徴とするATMスイッチ。

【請求項9】 請求項7において、上記複数個の入力バッファ内の各セルに対してキュータイムスタンプを付与するコネクション受付制御部と、

現在時刻を与えるタイマとをさらに備え、

上記入力モジュールスケジューラは、さらに、上記キュータイムスタンプと上記現在時刻を比較し、現在時刻に等しいキュータイムスタンプを有する上記セルの各々をスケジューリングするコンパレータを含むことを特徴とするATMスイッチ。

【請求項10】 請求項7において、上記コアスイッチの上記複数個の出力バッファとして、リアルタイム出力バッファが設けられると共に、上記コアスイッチは、更に、

上記リアルタイム出力バッファの数に対応した数だけ設けられ、上記出力ポートの一つにそれぞれ接続された複数個のノンリアルタイム出力バッファと、

上記出力ポートの各々に接続されたノンリアルタイムマルチキャストバッファとを備えたことを特徴とするATMスイッチ。

【請求項11】 請求項9において、上記入力モジュールスケジューラ及び上記出力モジュールスケジューラの各々は、

上記キュータイムスタンプを記憶し、上記コンパレータに上記キュータイムスタンプを提供するタイムスタンプ記憶装置と、

サービス資格のあるセルの伝送レートをシェーピングするための、上記コンパレータに接続された仮想レートシェーピング装置と、

サービス資格のあるセルの数をカウントするための、複数個の仮想キューカウンタと、

サービス資格のあるセルをスケジューリングするためのサービススケジューリング装置と、

上記タイムスタンプを動的に更新するための計算エンジンとをさらに備えたことを特徴とするATMスイッチ。

【請求項12】 請求項11において、記載の上記出力モジュールから上記入力モジュールへのレートフィードバックをさらに備えたことを特徴とするATMスイッチ。

【請求項13】 複数個の入力ポート及び複数個の出力ポートに接続された中央バスを有するコアスイッチと、上記出力ポートに接続された複数個の出力モジュールと、

上記入力ポートに接続された複数個の入力モジュールとを備え、当該各入力モジュールは、複数個の共通のキューと、スケジューラ部と、コネクション受付制御部とを

## 4

備えた入力モジュールとを含み、

上記コネクション受付制御部は、異なるクラスに上記キューを割り当て、入力セルストリームを、上記セルストリームに対して要求されたサービス品質にしたがって、適切なキューにルーティングすることを特徴とする、異なるサービス品質要求を有する異なるクラスのセルストリームをサポートできるATMスイッチ。

【請求項14】 コアスイッチと、上記コアスイッチの入力側に接続された複数個の入力モジュールとを備え、

10 上記コアスイッチは、

TDMバスと、

上記TDMバスに接続された複数個の入力ポートと、

上記TDMバスに接続された複数個の出力バッファと、

上記出力バッファにそれぞれ接続された複数個の出力ポートとを備え、

上記入力モジュールの各々は、

複数個の入力バッファと、

上記入力バッファの各々に対して最低保証レート及び超過配分加重を割り当てるコネクション受付制御部と、

20 上記最低保証レート及び利用可能な未使用帯域幅の割合から構成されるレートにしたがって、上記入力バッファ内のセルをスケジューリングするための入力モジュールスケジューラとを含み、上記配分は過剰配分重みに比例することを特徴とする、異なるサービス品質要求を有する異なるクラスのストリームをサポートできるATMスイッチ。

【請求項15】 請求項14において、上記出力バッファのセル占有率にしたがって、利用可能な未使用帯域幅を調整するための閉ループ制御器をさらに備えたことを特徴とするATMスイッチ。

30

【請求項16】 請求項15において、記載の上記出力バッファのうちのどれか1つのバッファのセル占有率が所定のシェーピングしきい値に到達するたびに、上記出力バッファのいずれかからシェーピングレート信号を送出するオーバーロード制御部をさらに備え、上記スケジューラは上記シェーピング信号を受け、保証された最小レートのみにしたがって上記入力バッファをスケジューリングすることを特徴とするATMスイッチ。

40

【請求項17】 入力セルストリームのキューを作るための複数個の入力バッファと、

上記入力バッファから上記セルストリームのキューセルを受取り、上記キューセルを各出力ポートに与える複数個のコアバッファと、

上記出力ポートから上記キューセルを受取り、上記キューセルを各出力回線に与える複数個の出力バッファと、上記入力バッファからの上記キューセルの送出をスケジューリングするためのスケジューラと、

上記スケジューラに上記コアバッファの負荷情報を伝送するための、上記コアバッファから上記スケジューラへの第1のフィードバックループと、

50

## 5

上記スケジューラに上記出力バッファの負荷情報を伝送するための、上記出力バッファから上記スケジューラへの第2のフィードバックループとを有し、

上記スケジューラは、上記第1及び上記第2のフィードバックループから受けとった情報にしたがって、上記入力モジュールからの上記キューセルの送出をスケジューリングすることを特徴とする、異なるサービス品質要求を有する異なるクラスのストリームをサポートできる、3重バッファ付ATMスイッチ。

【請求項18】 請求項17において、上記セルストリームに最低保証レートを割り当てるためのコネクション受付制御部をさらに備え、上記スケジューラは、上記最低保証レート以上のレートで、上記入力バッファにおいて上記キューセルをスケジューリングすることを特徴とする3重バッファ付ATMスイッチ。

【請求項19】 請求項18において、上記スケジューラは、動的レートにしたがって上記入力バッファにおいて上記キューセルをスケジューリングし、上記動的レートは、上記最低保証レートに、上記第1及び上記第2のフィードバックループの情報から決定された未使用帯域幅の割当を加えてなることを特徴とする3重バッファ付ATMスイッチ。

【請求項20】 請求項19において、上記コネクション受付制御部はさらに、上記キューの各々にキュータイムスタンプを割り当て、上記スケジューラは、上記キュータイムスタンプと現在時刻を比較し、現在時刻以下のキュータイムスタンプを有する各キューをサービス資格ありと定義するコンパレータをさらに備えたことを特徴とする3重バッファ付ATMスイッチ。

【請求項21】 スケジューリングすべきキュー内のセルに割り当てられたセルタイムスタンプを記憶するための第1のメモリと、

実際のキュー負荷を記憶するための第2のメモリと、  
仮想キュー内のセルを記憶するための第3のメモリと、  
現在時刻を発生する現在時刻発生器と、

上記セルタイムスタンプと現在時刻を比較し、現在時刻以下のセルタイムスタンプを有するセルをサービス資格ありと定めるための複数個のコンパレータと、

上記サービス資格のあるセルから上記仮想キューにセルを割り当てるための仮想レートセクタと、

上記仮想キューからサービス対象のキューを選択するサービススケジューリングセクタと、

上記第1メモリ内のセルをスケジューリング及び再スケジューリングするための計算エンジンとを備えたATMスイッチのためのスケジューラ。

【請求項22】 バッファであって、上記バッファにおける負荷レベルを監視し、上記負荷レベルが第1のしきい値に到達すると、上記バッファへの入力を最小レベルに低減するようにシェーピング信号を発生し、さらに、上記負荷レベルが第2のしきい値に到達すると、上記バ

## 6

ッファへのあらゆる入力を停止するようにストップ信号を発生するための第1の監視回路を有するバッファ。

【請求項23】 請求項14において、上記バッファ上で利用可能な未使用帯域幅を推定し、上記推定を示す信号を発生するための推定回路をさらに備えたバッファ。

【請求項24】 スケジューリングすべきキューに割り当てられたキュータイムスタンプを記憶するための第1のメモリと、

実際のキュー負荷を記憶するための第2のメモリと、

10 仮想キュー内のセルを記憶するための第3のメモリと、

現在時刻を発生する現在時刻発生器と、

上記キュータイムスタンプと現在時刻を比較し、現在タイムスタンプ以下のセルタイムスタンプを有するキューをサービス資格ありと定めるための複数個のコンパレータと、

上記サービス資格のあるセルから上記仮想キューにセルを割り当てるための仮想レートセクタと、

上記仮想キューからサービス対象のキューを選択するサービススケジューリングセクタと、

20 上記第1メモリ内のセルをスケジューリング及び再スケジューリングするための計算エンジンとを備えたATMスイッチのためのスケジューラ。

【請求項25】 互いにレートの異なるデータを伝送する複数のクラスのセルストリームをスイッチするATMスイッチにおいて、入力側及び出力側を備えたコアスイッチと、前記コアスイッチの入力側に接続された少なくとも一つの入力モジュールと、前記コアスイッチの出力側及び出力回線に接続された少なくとも一つの出力モジュールとを備え、前記出力モジュールから前記入力モジュールに対して、前記出力回線上の送信レートをフィードバックする少なくとも一つのフィードバックループが設けられていることを特徴とするATMスイッチ。

【請求項26】 請求項25において、前記入力及び前記出力モジュールには、それぞれスケジューラが設けられており、前記各スケジューラでは、前記各クラスのセルストリームの最低保証レートが、維持されるように、スケジューリングすることを特徴とするATMスイッチ。

【請求項27】 請求項26において、前記各スケジューラは、前記最低保証レートに加えて、各クラスの優先度に応じて定められた割合で、超過レートを割り当てることを特徴とするATMスイッチ。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、非同期転送モード(ATM)ネットワークに関し、特に、ATMフォーラムにより定義されている種々のクラスにおけるソースからの要求に効率的に応え得る大容量マルチクラスコアATMスイッチに関する。

【0002】

50 【従来の技術】過去、電話ネットワークとコンピュータ

ネットワークは、異なる分野で開発されてきた。効果的なリアルタイム通信を保証するために、電話TDM(Time Division Multiplexing)ネットワークでは、呼が継続している間維持されるチャンネルが設定される。一方、コンピュータネットワーク上で転送されるデータの多くはリアルタイムデータではないため、チャンネルを設定することなく、パケットをルーティングするパケット交換が行われている。

【0003】TDMネットワークにおける一つの問題点は、不使用状態にあるソースが設定されたチャンネルを不必要に占有してしまうことである。他方、パケット交換に伴う問題点は、パケット交換では高いプロトコルオーバーヘッドを必要とするため、リアルタイム通信に適さないということである。

【0004】非同期転送モード(ATM)技術は、将来の通信交換および伝送インフラストラクチャのための重要な技術として登場してきた。尚、ATMネットワークに関する記事を集めたものとして、「Lecture Notes in Computer Science」(Broadband Network Teletraffic, James Roberts, Ugo Mocchi, Jorma Virtamo (編者), Vol. 1155, Springer 1991年, ISBN 3-540-61815-5)がある。

【0005】ATMネットワークの主な長所は、互いに異なる種々のトラフィック特性及びサービス品質(QoS)要求をとともなうアプリケーションをサポートできる点にある。ATMネットワークの目標は、TDMネットワークとパケット交換の長所を組み合わせるとともに、これらネットワークのそれぞれの短所を取り除くことである。すなわち、ATM交換は、TDMネットワーク及びパケット交換ネットワークに取って代わるネットワークを提供することができる。

【0006】ATMフォーラムはATM設計のために種々のガイドラインを制定してきた。これらのガイドラインは、ATMフォーラムの各種刊行物において見ることができる。ここでは、説明の便宜上、関連するいくつかの指針及び略号を以下に説明する。

【0007】現在、ATMフォーラムは、4つの主要なトラフィックサービスクラス、即ち、固定ビットレート(CBR)、可変ビットレート(VBR)、利用可能なビットレート(ABR)、及び無指定ビットレート(UBR)のサービスクラスを定めており、これらは、リアルタイムトラフィック及びノンリアルタイムトラフィックのためのサービスクラスに分類される。CBRサービスはリアルタイムトラフィックに使用され、特に、音声トラフィックに用いられる。CBR呼要求の受付の可否は、要求されたピークレートにより決定することができる。VBRサービスは映像伝送に使用される。このサービスは非常にバースト性が高いトラフィックを対象とし

ており、VBR呼要求の受付の際には、ピークレートと、持続レートと、バーストサイズとが考慮される。受付けた際にも、リーキパケットを用いるなどして、このようなソースからの伝送を調整することが望ましい。上記したCBR及びVBRサービスは、いずれも、リアルタイムトラフィックを処理する。

【0008】ABR及びUBRサービスは、ノンリアルタイムトラフィックを扱うサービスであり、主に、コンピュータ通信に用いられる。従来、ABRトラフィックは、閉ループフィードバックを用いて制御され、これは、約3%のオーバーヘッドを考慮している。

【0009】一般に、ソースは、ネットワーク中を伝搬するリソース管理セル(RMセル)を生成する。各RMセルがスイッチを通過する際、サポート可能なレート、すなわち、ソースがデータを送出すべきレート(一般に明示レートと呼ばれる)を示すように、RMセルは更新される。これらのRMセルはソースにフィードバックされ、ソースはその伝送レートをRMセルの明示レートに応じて調整することができる。但し、このようなフィードバックシステムはかなりの遅延を有し、したがって、リアルタイムトラフィックには用いることができない。

【0010】一方、サービスクラスに応じて、ソースは適切なサービス品質(QoS)を要求する。一般に、QoSは、伝送遅延と、セル損失率と、セル損失遅延変動に基づいて決定される。上述のように、呼を受け付けた場合でも、例えば、リーキパケットを用いてピークレートを制御することにより、ソースにおける伝送を調整することができる。したがって、コネクション設定において、ソースは、適切な使用量パラメータ制御(UPC)値を交渉し、望ましいQoSを申告する。次に、コネクション受付制御(CAC)により、ネットワークが呼をサポートできるかどうかを決定する。

【0011】ソースは、さらに、宛先アドレスも指示する。宛先アドレスを用いて、ATMネットワークは、仮想チャンネル(VC)を設定し、適切なVCインディケータをソースに与える。ソースは、各伝送セル内にVCインディケータを挿入する。チャンネルは、呼の継続中は、不変である。すなわち、呼の全てのセルは、同一チャンネルを経由してルーティングされる。しかしながら、当該チャンネルは他のソースと共有されることもあるため、仮想チャンネルと呼ばれる。このことは、チャンネルとソースの間には、一対一の対応関係がないことを意味している。

【0012】一般に、受け付けた呼は、ATMスイッチ内のあるバッファに対応づけられ、スケジューリングアルゴリズムによって、いずれのバッファ、すなわち、いずれの呼が目的の時間に処理されるかを決定する。スケジューリングは、好ましくは、呼受付時に保証されたQoSを考慮し、ネットワーク資源によって公平に共有されるべきである。また、このスケジューリングにおける

アルゴリズムは動作維持性を有していること、すなわち、セルがバッファ内に存在している場合には、停止すべきではないことが望ましいとされている。

【0013】この種、ATMスイッチを含むATM交換システムが、英国特許第2,272,820号(以下、引用例と呼ぶ)において提案されている。このATM交換システムは、データトラフィックがピークの状態にある期間においても、ATMスイッチ動作を容易に行うことができるATM交換システムが開示されている。この引用例に示されたシステムは、複数の入力ポート及び複数の出力ポートを備えたATMスイッチ、当該入力ポートに接続された入力ポートサーバ、及び前記出力ポートに接続された出力ポートサーバとを有しており、各入力ポートサーバには、複数のバッファが設けられている。このシステムでは、データを各バッファから出力ポートサーバに送信する前に、各バッファから通信すべき出力ポートサーバに帯域幅要求を出し、出力ポートサーバにおいて、処理可能な容量があるかどうかが決定されている。容量が或る場合には、ATMスイッチ動作が行われ、他方、容量がない場合には、キューイングが行われている。

【0014】更に、ATMネットワークに使用できるスイッチとしても、種々のスイッチが提案されている。これら提案されたスイッチは、一段構成あるいは小さい一段スイッチを多段に接続した多段構成を有している。スイッチは、一般に、セルバッファの位置によって、入力バッファ型及び出力バッファ型かに分類できる。出力バッファを設けることによって最適なスループットが達成されることは良く知られている(例えば、M. J. Karol, M. G. Hluchy, S. P. Morganによる「Input vs. Output Queueing on a Space-Division Packet Switch」(IEEE Trans. Comm., Vol. 35, 1347~1356頁、1987年12月)参照)。

【0015】

【発明が解決しようとする課題】上記した引用例に示されたATM交換ネットワークでは、出力ポートサーバに、帯域幅を受けられる容量があるか否かが決定されている。より具体的に言えば、入力ポートに負荷が増えてきた場合、空き帯域があるかどうかを問い合わせ、ある場合には、帯域を予約しておく。逆に、負荷が減ってきた場合、予約している帯域を開放しなければ、他のクラス/ポートがその帯域を利用できないため、その帯域が無駄になってしまう。また、実際、負荷の変化は、非常に激しく、且つ、予想しにくいいため、引用例に示されたネットワークにおいては、トラフィックの変化に追随できず、このため、帯域の利用率が悪くなってしまう。

【0016】更に、帯域幅要求があった場合に、帯域幅自体を変化させることについて、引用例では、何等考慮

されていない。更に、この引用例においては、複数のサービスクラスに適用することについても、全く開示していない。

【0017】一方、出力バッファを有するスイッチ構成では、出力バッファは、ラインレートのN倍のアクセス速度で動作することが必要である。ここで、Nは入力ポートの数である。スピードアップ係数Nは、いわゆる「ノックアウト原理」(Y. S. Yeh, M. G. Hluchy, A. S. Acamporaによる「The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching」(IEEE J. Select. Areas Comm. Vol. 5, 1274~1283頁、1987年10月)参照)を用いることにより、N=8まで下げることができる。しかしながら、スイッチに不均一なトラフィックパターンが加えられた場合には、望ましくないセル損失が生じることがある。共有メモリを有するスイッチにおいても、共有メモリはN倍アクセス速度を持たなければならない。

【0018】入力バッファを備えたスイッチはスピードアップを一切必要としないが、ヘッドオブライン(HOL)ブロッキングによるスループット低下が生じる。すなわち、宛先出力回線が先頭ラインセルを受け付ける準備ができるまで、入力バッファキューの先頭のセルがバッファ内の他の全てのセルをブロックする。しかしながら、先頭ラインセルによりブロックされている他のセルを他の宛先出力回線では実際には受け付けることができる場合もある。このように、受付可能なセルを不所望にブロックすることは帯域幅の効率的な使用を妨げ、不必要な遅延を引き起こすことになる。

【0019】現在のATMスイッチは比較的単純なスケジューリングと、QoSに対するサポートが限定されたバッファ管理メカニズムを有する。一方、ATM技術が広域網(WAN)市場に普及するにつれて、より多様なアプリケーションから、かなり大量のトラフィックを扱うWANコアでは、より高度なスイッチが必要となる。次世代WANコアスイッチは、大容量と、マルチクラスのトラフィックに対してQoSサポートを行なえる能力を備えることが必要であると考えられる。

【0020】したがって、本発明の目的は、このような多様なトラフィックをサポートできるATMスイッチを提供することである。

【0021】本発明の他の目的は、帯域の変化に容易に対応出来るATMスイッチを提供することである。

【0022】本発明の他の目的は、帯域幅を変化させることができるATMスイッチを提供することである。

【0023】

【課題を解決するための手段】本発明によれば、複数のトラフィッククラスとサービス品質(QoS)保証をサ



ポートする大容量ATMスイッチが得られる。このスイッチは、CBRやVBRなどのきびしいQoS要求をともなうリアルタイムトラフィッククラスと、ABRやUBRなどのそれほどきびしくない要求をともなうノンリアルタイムトラフィッククラスの両方をサポートする。この構成は、さらに、リアルタイム及びノンリアルタイムのマルチキャストフローを効率的に調整できる。スイッチは、大容量バッファを備えた入力／出力モジュールを相互接続する高速コアスイッチモジュールを備えると共に、入力及び出力バッファの双方を有している。この場合、コアスイッチモジュールには、小容量の高速のコアバッファが設けられる。このように、入力バッファ、コアバッファ、及び、出力バッファの3バッファ構成が採用され、バックプレッシャー、即ち、出力バッファから入力バッファへのフィードバックを掛けることにより、入力ポート間の制御を避けることができ、且つ、大容量の出力バッファの速度を低下させることができる。尚、入力バッファは、出力ポートに対応して設置されても良いし、出力ポート並びに出力回線に対応して設置されても良い。入力バッファを出力ポートに対応して設けた場合、出力ポートへのトラヒックをスケジューラで制御できるが、出力回線の区別がないため、大きな出力バッファが必要となる。一方、入力バッファを出力ポート並びに出力回線に対応して設けた場合、出力回線へのトラヒックを制御できるため、出力バッファの容量を減少させることができる。

【0024】このように、インテリジェントスケジューリング及びキュー管理メカニズムによって、コアスイッチモジュールは制御されたクラスにしたがってアクセスされる。

【0025】スイッチは、好ましくは、本発明者らにより発明され、同日付で提出される新規なスケジューリング方法、即ち、動的レート制御（DRC）とともに用いられる。このDRCは、内部輻輳を制御し、スイッチのボトルネック部で、競合するフロー間の公平なスループット性能を達成する。これは、各ボトルネック部において比例微分[proportional-derivative]（PD）制御器を用いた閉ループ制御を行なうことにより達成される。DRC方式は最低保証レートに、未使用帯域幅を公平に配分する動的レート成分を加えたレートを、各フローに保証する。このことは、異なるトラフィックサービスクラスに対するQoSを統合したシステムにおける基盤を与えることができる。

【0026】本発明に係わる大容量スイッチでは、DRCスケジューリングメカニズムは、インテリジェントキュー管理メカニズムとともに動作する。DRCスケジューラは、スイッチ内のボトルネック部において輻輳を検出し、スイッチの入力側に、セルを移動させることにより、輻輳を緩和する。ここで、初期パケット廃棄（EPD）や部分パケット廃棄（PPD）などのセル廃棄メカ

ニズムを、個々のクラスキューに適用してもよい。また、低い優先順位をタグ付けされたセル、すなわち、セル損失優先（CLP）ビットが1に設定されたセルは、キューがしきい値を越えると、廃棄される。DRCを用いると、スイッチ輻輳レベルにしたがって制御され、セルが廃棄されるため、セル廃棄メカニズムはより効果的に動作する。

【0027】ここに述べる本発明の大容量スイッチは、総スループット及びマルチクラスQoSに対するサポートの両方において、現在のスイッチに対して著しく改善されている。また、本発明に係わる大容量スイッチはフレキシビリティとスケラビリティを有しており、現在及び将来の高性能ATMネットワークの要望を満足させることができる。

【0028】

【発明の実施の形態】

#### 1. 全般的な構成

好ましい実施例において、本発明の大容量スイッチは一段スイッチであり、この一段スイッチは、より大容量の多段スイッチにおけるスイッチング素子として用いることもできる。本発明による大容量スイッチは、入出力バッファを備えたスイッチ（以下、入出力バッファ付スイッチと呼ぶ）（R. Fan, H. Suzuki, K. Yamada, N. Matsuuraによる「Expandable ATOM Switch Architecture (XATOM) for ATM LANs」(Proc. ICC '94, 99~103頁, 1994年5月)参照)として分類できる。入出力バッファ付構造の目標は、入力バッファ及び出力バッファの長所を組み合わせることである。

【0029】好ましい実施例において、出力バッファは、コアスイッチモジュールの一部である小型で高速のバッファである。セルはライン速度で動作するバッファを有する入力モジュール内で主にキューイングされる。宛先出力ポート、あるいは、宛先出力回線にしたがって、入力モジュール内のセルをキューイングすることによりHOL（ヘッド・オブ・ライン）ブロッキングは回避される。この構成は、高速で大容量の出力バッファを必要とすることなく、出力バッファ付きを用いた場合と同等のスループットを達成できる。その上、入力ポートにバッファを設けることは、出力バッファを設ける場合よりも有効である。セル損失性能を同じにした場合には、出力ポートにおいてキューイングするよりも入力ポートにおいてセルのキューイングを行った方が、全体としてはバッファの容量を少なくできる。

【0030】本発明は、高速且つ簡単であり、クラスに関係なく、しかも、無損失でスイッチングを行なえる大容量マルチクラススイッチを提供することを主目的としており、このスイッチは、新規な高速コアスイッチ素子によって構成されている。コアスイッチモジュールに対

するアクセスは、入力モジュール (IM) のインテリジェントスケジューリングメカニズムにより制御される。入力モジュールは、クラス毎のキューイング、或いは仮想チャネル毎のキューイングができるように構成されている。

【0031】大容量スイッチの好ましい実施例の全体図を図1に示す。図示の実施例において、コアスイッチモジュール10は、16個の入力ポートIP1~IP16と、16個のユニキャスト出力ポートOP1~OP16と、1個のマルチキャスト出力ポートMOPによって構成され、これら全ては、高速TDMバス20に接続されている。図示の実施例において、コアモジュールの入出力ポートは2.4Gbpsのレートで動作し、TDMバスは40Gbpsのレートで動作する。この例では、1セル時間内に、コアスイッチモジュールは、各入力ポートからの一つのセルを出力ポートのいずれかに交換することができるものとする。

【0032】入力モジュール (IMi) は、コアモジュールの各入力ポートIPiに接続される。ここで、入力モジュールIMiの出力回線容量は2.4Gbpsであるものとする。入力モジュールIMiの入力側は入力回線IL1i~IL16iに接続され、これらの入力回線は次の3通りのいずれかの構成を有している。すなわち、1) 1本の2.4Gbps回線、2) 4本の622Mbps回線、3) 16本の155Mbps回線のいずれかによって構成される。図示された実施例のすべての場合においても、入力モジュールIMiの総入力回線容量は2.4Gbpsである。入力回線IL1i~IL16iには、異なるQoS要求をとまう異なるクラスに分類されるソースから、伝送が行なわれる。

【0033】仮想チャネル (VC) 毎のキューイング (図2a) と、出力ポートに対するクラス毎のキューイング (図2b) と、出力回線に対するクラス毎のキューイング (図2c) に応じて、入力モジュールは構成される。VC毎のキューイングは、スループットの点では最良であるが、各入力モジュールに多数のバッファを必要とする。したがって、VC毎のキューイングは設計上では望ましいが、実現性の面では好ましくない。

【0034】図2bに示すように、出力ポートに応じたクラス毎のキューイングにおいて、各入力モジュールIMiは、出力モジュールの数に対応して、多数のレイヤ/プレーンLOP1~LOP16、本実施例では、16個のレイヤを有している。各レイヤは、サポートする必要があるクラス数kに対応して数個のバッファIB1~IBkを備えている。また、各入力バッファIBiには、サービスレートRi1~Rikが保証されている。これにより、入力セルは、宛先出力ポートに対応する適正なレイヤに、そして、そのクラスに応じたレイヤ内の適切な入力バッファにルーティングされる。出力ポート上の負荷に対応して、DRCレートが出力モジュールから

らフィードバックされる。

【0035】図2cに示すように、出力回線に応じて、クラス毎のキューイングを行なうためにスイッチが構成されている場合、入力モジュール内には2組のレイヤ/プレーンが設けられている。まず、出力モジュールは出力ポートに対応する出力ポートプレーンOP1~OP16に分割されている。次に、各出力ポートプレーンを出力回線に対応する出力回線プレーンOL1~OLkに分割されている。最後に、各出力回線プレーンは各クラスに対応する複数のバッファを備えている。この場合、2個のDRCレートがフィードバックされ、一つは、出力ポート上の負荷、他方は、出力回線上の負荷をそれぞれ示している。

【0036】同様に、出力モジュール (OMi) を各ユニキャスト出力ポートOPiに接続する。OMの出力側は、2.4Gbpsの総出力回線容量を有する1本、4本、あるいは、16本の出力回線からなる。出力モジュールは、出力回線に対応する出力プレーンに分割される。各出力プレーンは、サポート可能なクラスに対応する複数のバッファを有する。

【0037】コアスイッチモジュール10の各出力ポート (OPi) は、2個の小さい出力バッファRTi、NRTiに組み合わせられ、一方のRTiは、リアルタイムトラフィック用に指定され、他方のNRTiはノンリアルタイムトラフィック用に指定されている。好ましい実施例において、各出力バッファRTiあるいはNRTiは、200個程度のセルを記憶することができる。各ユニキャスト出力ポートの出力において、ノンリアルタイム及びリアルタイムバッファ用出力回線は、マルチキャスト出力ポートの対応する出力回線と組み合わせられる。各セル時間のあいだ、単一のセルがユニキャスト出力ポートOPiの出力から、対応する出力モジュールOMiに伝送される。

【0038】図1を参照すると、優先順位は以下の

1)、2)、3)、4)の順で低くなる。1) マルチキャストリアルタイムトラフィック、2) ユニキャストリアルタイムトラフィック、3) マルチキャストノンリアルタイムトラフィック、4) ユニキャストノンリアルタイムトラフィック。

【0039】コアスイッチモジュールがユニキャスト及びマルチキャスト交換を実行するための基本ハードウェアを提供するが、好ましい実施例において、当該スイッチの本発明の対象となる部分は、実質上、入力モジュール (IMi) と出力モジュール (OMi) にある。各入力/出力モジュールには、スケジューラと、キュー管理部、即ち、マネージャーと、大容量のスペースを有するバッファが設けられている。コネクション受付制御器と入力/出力モジュール間とは、密に結合されているため、キューフローがそのサービス品質 (QoS) 要求を確実に満足させることができる。

【0040】各入力モジュールIMは、ATMセルヘッダ変換動作を行なうと共に、VC(図2a)及びクラスとコアスイッチ宛先出力ポート(OPi)(図2b)、あるいは、クラスと宛先出力回線(図2c)により構成されるキュー内の入力セルが格納される。ここで、キューフローの用語は、与えられたキューに対応する全てのコネクションの総トラフィックをあらわすために用いられる。コネクション受付制御器(CAC)(図3参照)はどのトラフィッククラスのキューに対しても、フレキシブルにQoSを割り当てることができ、ここでは、このQoSをプログラマブルQoSと呼ぶ。

【0041】一方、クラスキューは、リアルタイムあるいはノンリアルタイムキューに分類される。

【0042】各セル時間中、入力モジュールIMi内にキューがあれば、スケジューラがキューのうちの一つを選択する。選択されたキューから、ヘッドオブライン

(HOL)セルがTDMバスを経由して宛先出力ポートOPiに伝送される。キューマネージャーはキューにセルバッファを割り当て、バッファしきい値を超過した場合には、セルを廃棄する。入力モジュールIMi内のセルのキューイングは、コアスイッチモジュール内の出力ポートにおける輻輳を回避するように設計されている。異なる入力モジュールIMiからのキューフローの合計が出力ポートOPiにおける容量Cを越える場合に、輻輳が生じることもある。このような状況のもとで、出力ポートOPはボトルネック部となる。

【0043】出力モジュールの構成は、入力モジュール構成と同様である。各出力モジュールは、基本的にスイッチの残りの部分から独立して、且つ、判断機能を備えたインテリジェント多重分離器として動作する。出力モジュールOMi内で、セルはクラス及び宛先出力回線にしたがってキューイングされる。出力モジュールOMiにおけるセルのキューイングは、出力モジュールOMiに接続された出力回線OLiにおける輻輳によって生じる。すなわち、出力回線に対するキューフローの合計がその容量を越えた場合に、出力モジュールにおけるセルキューが生じる。このように、出力回線OLiは、レートの不適合が生じた場合、スイッチ内のキューフローに対する他のボトルネック部となる。出力ポートOPi及び出力回線OLiのようなボトルネック部における内部スイッチ輻輳は、入力/出力モジュール内のインテリジェントスケジューリングとキュー管理によって制御される。

【0044】図3は、入力及び出力モジュールの構成をより詳細に示している。図示された入力モジュール30は、16個の出力ポートOP1~OP16に対応する16個のプレーンを備えている。各プレーンは複数の(4個のみ図示されているが、より多くてもよい)同一構成のバッファ32を備えている。これらのバッファ32はCAC33により、それぞれのQoS要求を有する

多様なクラスに対応するようにプログラムされている。図示された入力モジュール30において、4個のバッファはCBR、VBR、ABR、UBRの4つのサービスクラスにそれぞれに割り当てられる。CAC33は、さらに、バッファ内のセルに対して、送受信時点を指示するタイムスタンプを与える。

【0045】コアスイッチモジュール34は、複数のバッファ付き出力ポートiに接続されたTDMバス35を備えている。図示の例では、各出力ポートは2個のバッファ、すなわち、リアルタイムバッファRtとノンリアルタイムバッファNr tを有し、出力バッファは出力モジュールに接続されている。各出力モジュールは出力回線に対応するプレーンOL1~OLkに分割されている。さらに、これらの出力プレーンには、複数のプログラマブルバッファ及び1個のスケジューラが備えられている。

#### 【0046】2. マルチキャストイング

本発明によるスイッチの特徴は、ユニキャストだけでなくマルチキャストを効率的にサポートできる点にある。

スイッチ内のユニキャストコネクションは、入力モジュールIMiへの入力回線から、出力モジュールOMiの単一の出力回線までのコネクションである。一方、マルチキャストコネクションは、入力モジュールIMiから、一つ以上の出力モジュールOMiにおける複数の出力宛先回線との間のコネクションである。コアスイッチモジュール34は、複数の出力モジュールOMi間のマルチキャストをサポートできる。2つ以上の出力モジュールOMiにマルチキャストされるセルは、入力モジュールIMiからマルチキャスト出力ポートMOPに送られる。

【0047】当該マルチキャストセルも、リアルタイムあるいはノンリアルタイムセルに区分され、当該セルのコピーが一つだけマルチキャスト出力ポートMOP内の対応するバッファに格納される。出力モジュールOMiに送出される直前に、マルチキャスト出力ポートMOPの出力側でセルが複写される。

【0048】図1に関連して述べられたように、リアルタイムマルチキャストトラフィックは各出力モジュールOMiに入力される場合、最優先順位を有している(図1において、出力における優先順位は、出力モジュールに対する入力をあらわし、且つ、垂直線方向に向いた矢印の順番により示されている。したがって、垂直線に対して一番上にある矢印が最優先順位を示している。)

【0049】与えられたセル時間内に、マルチキャスト出力ポート用リアルタイムバッファの先頭にマルチキャストセルがある場合、セルは複写され、マルチキャストに関与する出力モジュールOMiに送られる。ここで、マルチキャストセルの複写は、コアスイッチモジュールのTDMバス上では行われない。

【0050】リアルタイムマルチキャストトラフィック

は、最優先順位を有するため、ブロックされることはない。一方、ノンリアルタイムマルチキャストトラフィックは、リアルタイムマルチキャスト及びリアルタイムユニキャストよりも低優先順位を有している。

【0051】このことを考慮すると、与えられたセル時間内において、出力ポートOPiにリアルタイムマルチキャストセルが或る場合、或いは、マルチキャストの対象となるユニキャスト出力ポートOPiのいずれかにユニキャストリアルタイムセルがある場合には、出力ポートOPiにおけるノンリアルタイムマルチキャストセルはブロックされることになる。

### 【0052】3. フィードバック制御

フィードバック制御は、スイッチの効率的な動作を保証するために用いられる。コアスイッチモジュール内の出力ポートバッファは小規模であるため、すぐにオーバーフローする可能性がある。そこで、このようなオーバーフローを調整するために、2つの基本的なフィードバックメカニズムが用いられる。これらのフィードバックメカニズムには以下のようなメカニズムがある。

【0053】1. 出力ポートバッファにおけるキューを短く保ちながら、ボトルネックレートに合致させ、且つ、利用度を高く維持した閉ループフィードバックメカニズム（第一の制御メカニズム）。

【0054】2. コアスイッチモジュールの出力ポートバッファが、第一の制御メカニズムによる制御にもかかわらずオーバーフローする可能性を有する場合に起動されるしきい値ベースのレートフィードバックメカニズム（第二の制御メカニズム）。

【0055】第一の制御は、入力モジュールにおける動的レート制御（DRC）スケジューリングにより達成される。第二の制御は、コアスイッチモジュール内に組み込まれ、出力ポートOPiボトルネックにおける短期的輻輳を迅速に制御するための安全メカニズムとみなされる。コアスイッチモジュールは、各セル時間中、全ての入力モジュールIMiに出力ポートOPiの状態情報を

同報的に送るためのフィードバック経路を有している。フィードバック信号が出力ポートから入力モジュールに伝搬する時間は、本明細書中に $\tau_d$ （セル時間）であらわされ、この量はシステムに応じて異なる。

【0056】好ましい実施例において、スケジューラは、入力モジュールに対して未使用帯域幅を配分する。この場合、入力モジュールからの実際の伝送レートは、保証最小レートを越えることもある。しかしながら、状況によっては、全ての利用可能な帯域幅を用いた場合、ある出力ポートで輻輳を生じることがある。そこで、好ましい実施例において、このような輻輳を少なくするためにフィードバック信号が使用される。

【0057】好ましくは、入力ポートバッファに関連して、3つのしきい値が、制御フィードバック信号として生成される。これらしきい値は、1) ストップリアルタイム（RT）、2) シェーピングリアルタイム（RT）、3) ストップリアルタイム（NRT）がある（図1参照）。ストップRTしきい値インディケータは、リアルタイムバッファにおける格納量がしきい値 $Th_{stop}$ 以上である場合、1に設定され、それ以外の場合には、ストップRTインディケータはゼロとなる。同様に、ノンリアルタイムキューの格納量が $Th_{stop}$ 以上である場合、ストップNRTは1となる。ストップ信号が入力モジュールIMに到達するまで、全ての入力モジュールIMが同一の出力ポートにセルを送出するという最悪の状態ではバッファオーバーフローが生じないように、ストップしきい値 $Th_{stop}$ が最大値として選択されている。シェーピングRTインディケータは、リアルタイムバッファがシェーピング用しきい値 $Th_{shape}$ （ $< Th_{stop}$ ）以上の場合、1に設定される。表1に、しきい値インディケータと2ビットに符号化された出力ポートに対する制御信号B0、B1との関係、入力モジュールにより実行される動作が示されている。

【0058】

【表1】

しきい値インディケータ			制御ビット		入力モジュールで 実行される動作
シェーピングRT	ストップRT	ストップNRT	B1	B0	
0	0	0	0	0	送信RT、停止NRT
0	0	1	0	1	送信RT、停止NRT
1	0	0	1	0	シェーピングRT、 送信NRT
1	1	1	1	1	停止RT、停止NRT

ストップRTインディケータが1に設定されると、適切なフィードバック信号が送出される。 $\tau_d$ セル時間の後、信号は全ての入力モジュールに与えられ、各入力モジュールは、対応する出力ポートにセルフロー（リアルタイム及びノンリアルタイムの両方）を制限する。ストップNRTインディケータは、ノンリアルタイムトラフィックに対して、同様に機能する。上記したフィードバ

ック信号により、出力ポートにおけるセル損失を回避することができる。ここで、ストップ信号は、入力モジュール内の入力のキューイングを発生させる。ストップ信号がない場合、入力モジュール内におけるセルのキューイングは生ぜず、結果として、出力ポートにおいてオーバーフローが生じて、セル損失が発生してしまう。

【0059】シェーピングRTインディケータは、キュー

フローに対して予め割り当てられた最低保証レートに基づき、リアルタイムトラフィックの輻輳を制御する手段として動作する。

【0060】ここで、ある出力ポートOPjから入力モジュールIMiに対してシェーピングRT信号が受信されると、出力ポートOPjに対応する全てのリアルタイムキューフローは、その最低保証レートにシェーピングされる。すなわち、リアルタイムキューは、利用できる未使用帯域幅に関係なく、その最低レートでスケジューリングされる。この動作により、リアルタイムキューフローに対する最低保証スループットを確保しながら、出力ポートOPjにおけるリアルタイムキューの格納量が大きくなるのを防ぐことができる。

【0061】このように、リアルタイムキューフローは、輻輳がない場合、最低保証レートより大きいレートで処理され、輻輳がある場合には、最低保証レートに等しいレートで処理される。

【0062】以下に詳述するように、DRC（動的レートコントロール）スケジューラでは、リアルタイムキューフローに対しては、ストップ信号が低い確率しか発生しない。

【0063】4. トラフィッククラス及びサービス品質(QoS) サポート

#### 4. 1 リアルタイムトラフィック

CBRやVBRなどのリアルタイムトラフィックでは、セル遅延と、セル損失と、セル遅延変動(CDV)に対して厳しい要求が加えられている。コネクション受付制御(CAC)アルゴリズムについて言えば、大容量スイッチ構造を採用することによりQoS保証を提供することができる。このようなCACアルゴリズムとして、

G. RamamurthyとQ. Renによる「Multi-Class Connection Admission Control Policy for High Speed ATM Switches」(proc. IEEE INFOCOM '97(神戸、日本)1997年4月)によって提案された手法を本発明に係わるスイッチのCACにおいても用いることができる。このCACは、リアルタイムキューフロー内のコネクション全てのQoS要求を満足させるように、必要な帯域幅を算出する。すなわち、CACは、任意のキューフロー内のコネクションの統計的多重化を考慮している(統計的多重化は、全ストリームに必要な帯域幅が各ストリームに必要な個々の帯域幅の合計よりも少ないということを考慮した手法である)。必要な帯域幅は、各コネクションに対するUPCパラメータ及びフローに対するバッファの予備割当て量に基づいて算出される。しかしながら、この従来技術では、計算された最低レートはCACの目的のためだけに用いられ、スケジューラでは使用されていない。

【0064】DRCスケジューリングメカニズムは、各

キューフローがその最低保証レートを受けとり、したがって、QoSがフロー内のコネクション全てについて確実に保証する。コアスイッチ素子においてリアルタイムトラフィックがノンリアルタイムトラフィックに対して高い優先権を持ち、出力ポートからのシェーピングフィードバックメカニズムにより、輻輳状態においても、当該キューフローの最低保証レートが保証されるから、最低レートは短時間内においても保証される。すなわち、輻輳状態のもとでは、シェーピングフィードバックメカニズムにより、未使用帯域幅の分配が停止され、これにより、レートを最低保証レートに低下させて輻輳を緩和する。この状態では、最低保証レートが確実に保証される。さらに、シェーピングモードにおいて、最低レートでの動作を行なっているキューでは、サービスを受けることができる場合、当該優先ビットがセットされる。

#### 4. 2 ノンリアルタイムトラフィック

ノンリアルタイムトラフィッククラスにはABR及びUBRが属している。これらのクラスは、通常、QoSに対する要求は厳しくないが、最低スループットが要求されることがある。ノンリアルタイムキューフローについての最低レートは、フロー内のコネクション全てについての最低スループットの合計に等しい。大容量スイッチスケジューラは、DRCスケジューリングを行なうことによって、各ノンリアルタイムフローに対して最低レートを保証できる。スイッチボトルネックにおける未使用帯域幅も、競合するキューフロー(リアルタイム及びノンリアルタイムの両方)間で分配される。未使用帯域幅の分配は、異なるトラフィッククラスに割り当てられた重み $w$ に依存して行なわれ、好ましくは、これらのレートは動的に割り当てられる。

【0065】UBRソースはレート制御されず、従来のスイッチ構造においてスループットの損失を引き起こし得る。動的レート制御では、UBRキューには、その最低レートに、未使用帯域幅の公平な割当てを加えたものが与えられる。ABRソースは閉ループフィードバックメカニズムによってレート制御される。スイッチにおいて、明示レート(ER)値が、コネクションフローにおける各ボトルネック部において算出される。大容量スイッチ構造において、ABR ER値は、出力モジュール内の出力ポートボトルネック部及び出力回線ボトルネック部において算出される。ERを算出するために様々な方法が用いられる。しかしながら、好ましい実施例において、ABR ER値はDRCレートの算出と同様に算出される。

#### 【0066】5. 動的レート制御

動的レート制御(DRC)は、大容量スイッチにおいて用いられるセルスケジューリングのためのメカニズムである。DRCについては米国特許出願番号08/92

4, 820号に記載されているが、説明の便宜上、本発明によるスイッチに適用した場合の、DRCの基本原則

の概要を以下に説明する。また、A. KolarovとG. Ramamurthyによる「Design of a Closed Loop Feed Back Control for ABR Service」(Proc. IEEE INFOCOM '97 (神戸、日本) 1997年4月)には、ATMネットワークに適用した場合のABRサービスに対するフィードバック制御が記載されている。しかしながら、以下の説明におけるフィードバックは、ATMスイッチに適用されていることは留意すべきである。

【0067】DRCの基本原理は、各クラスのキューを仮想ソースと同様に扱い、そのサービスレートを動的に調整することにより、スイッチ内のボトルネック部における利用可能な未使用帯域幅に反映させることである。すなわち、各クラスは、その最低保証レートに、利用可能な未使用帯域幅の動的に調整され、公平に配分されたレートを加えたレートで処理される。スケジューリングは、キューサービスレートを算出し、全てのキューについてレートシェーピングを行なうことによって実現される。スケジューリングに対するこの方法の重要な特徴は、全てのキューを平均的なキューの集合に変え、クラスのQoSを、そのクラスに対して保証された帯域幅によって決定することである。ここで、平均的なキューは、CACにより各クラスに割り当てられる。

#### 5. 1 最低保証レート

スイッチにおける任意のコネクションに対するQoSを保証するために、スイッチでは、トラフィック特性と帯域幅リソース間のマッピングが行なわれる。任意のコネクションiのトラフィック仕様には、セル損失確率と、遅延、および/または遅延ジッタであらわされた一組のQoS要求が含まれている。指定された要求の全てに関連したアルゴリズムを実現するよりも、これらの要求全てに関連した単一の変数にこれらの要求をマッピング、即ち、変換するほうが簡単である。

【0068】DRCの好ましい実施例では、要求が帯域幅、又は、レートMiにマッピングされている。具体的に言えば、コネクションiにレートMiを割り当てることにより、当該QoS要求が満たされることになる。好ましい例において、Miは、コネクション受付制御(CAC)アルゴリズムによって与えられる。この場合、レートMiは、QoSに対する要求の全てを満たすように、十分に近似されなければならない。

【0069】一旦、Miが決定されると、スケジューラはコネクションiに対して最低レートを保証する。このことは、コネクションiのQoSが保証されることを意味している。このように、上記したスケジューラはひとつの変数のみを考慮すれば良いから、構成によって簡単になる。

【0070】ここで、N個のコネクションが容量Cのリンク上に、多重化されるネットワーク内における集線位

置について考察してみよう。この場合、次の数1式の関係が成立することは明らかである。

【0071】

【数1】

$$\sum_{i=1}^N M_i \leq C$$

単純な先入れ先出し(FIFO)スケジューラを使用した場合、各コネクションに対して割当帯域幅Miを保証することはできない。例えば、あるコネクションがその割当分Miより高いレートで伝送し、これにより、別のコネクションから帯域幅を奪ってしまうことがある。

【0072】コネクションがその割当分より多くの帯域幅を使用しないように保証する単純な方法は、各コネクションiのピークレートをMiに限定することである。このことは、例えば、リーキバケットなどの公知の方法を用いて、割り当てられた最低レートに各コネクションのピークレートをシェーピングすることにより実施することができる。

【0073】図4は、各キューiを含むN個のキューがそれぞれのレートMi(i=1, ..., N)にシェーピングされた場合を示している。この例では、シェーピングされたトラフィックストリームは、次に多重化され、下流側バッファのレートC以下のレートでFIFO順に処理されている。

【0074】ピークレートによる制限によって、全てのコネクションについて最低レートを保証することができる。しかしながら、このスケジューリング規則では、コネクションiは、広い帯域幅が利用可能な場合でさえも、割り当てられたMiより多くの帯域幅を用いることはできない。例えば、コネクションiがリンク帯域幅Cを共用する唯一のアクティブなコネクションである場合、全リンク容量Cが利用可能であっても、帯域幅Miだけしか使用できない。さらに、統計的多重化が行なわれることを前提として最低レートMiを算出する場合、リンク容量を超える確率は小さいため、全帯域幅が共有された場合を仮定して、コネクションiのQoSが保証される。この種の共用は、コネクションのピークレートが値Miに制限される場合には起り得ない。統計的多重化利得を前提として算出される割り当て帯域幅Miは、統計的多重化が行われない場合には、QoSを保証するには不十分になるおそれがある。

【0075】一方、静的レートに基づくスケジューラのレート制御メカニズムは、開ループと呼ばれることがある。これは、スケジューラが動作維持型でないこと、すなわち、システム内で処理すべきセルがある場合でさえも、出力リンク上にセルが送出されない空白の時間があることである。この場合、利用可能な帯域幅がこれらのスケジューリング規則では、無駄になる可能性がある。

【0076】他方、上記したDRCスケジューラは、閉

ループ制御メカニズムによってこの問題を解決している。

【0077】DRCスケジューラの基本原理を図5に示す。前述のように、各トラフィックストリームは、リンクレートCの共通のFIFOキューに入力される前に、ピークレートについてシェーピングされる。しかしながら、シェーピングレート $R_i$ は、リンク上で利用可能な帯域幅の量を反映するように動的に算出される。具体的に言えば、コネクション $i$ は、 $R_i$ にピークレートシェーピングされるものとする、 $R_i$ は次式で与えられる。

$$【0078】 R_i = M_i + w_i E$$

ここで、 $E$ は、ボトルネック部における推定未使用帯域幅（本明細書中では超過レートあるいは超過帯域幅とも呼ぶ）である。一方、 $w_i \geq 0$ であり、静的あるいは動的に割り当てられる任意の重み係数である。通常、 $E \geq 0$ であるため、 $R_i \geq M_i$ が成り立つ。上式からも分かるように、コネクション $i$ には最低レート $M_i$ が保証されるが、未使用帯域幅が利用可能である場合には、より高いレートで伝送することもできる。逆に、輻輳中には、スケジューラが $E$ をゼロとすることにより、輻輳が解消されるまで、その最低保証レートでのみキューは処理されることがわかる。

## 5. 2 閉ループレート制御

図6は、DRCスケジューラにおいてコネクションに分配される利用可能な帯域幅を検出する閉ループレート制御システムを示している。ここで、時間を、長さ $T$ の区間に分割して、離散的にあらわすものとする。 $X_i$

( $n$ )は $n$ 番目の区間に、コネクション $i$ により発生されるセル数を示す。また、量 $Q(n)$ は第2段バッファ内のセル数を表わし、第1段バッファでは、各コネクションストリームは、次式で与えられるレートにしたがってシェーピングされるものとする。

【0079】

$$R_i(n) = \min(M_i + w_i E(n), C)$$

コントローラ、即ち、制御器は、 $Q(n)$ が目標のキューしきい値 $Q_0$ に近い値に保たれるように、 $E(n)$ を

$$\begin{aligned} E(n+1) = & E(n) - \alpha_0 \varepsilon(n) - \alpha_1 \varepsilon(n-1) - \dots \\ & - \alpha_u \varepsilon(n-u) - \beta_0 E(n) - \beta_1 E(n-1) - \dots \\ & - \beta_v E(n-v) \end{aligned} \quad (1)$$

ここで、 $\alpha_i$  ( $i=1, \dots, u$ ) と、 $\beta_i$  ( $i=1, \dots, v$ ) は、実数の係数である。

【0085】DRCスケジューラでは、単純な2パラメ

$$E(n+1) = E(n) - \alpha_0 \varepsilon(n) - \alpha_1 \varepsilon(n-1) \quad (2)$$

上記したことから、本発明に係るコントローラは簡略化した構成を備えると共に、レートシェーピングの速度を向上させることができる。

## 5. 3 オーバードロード（過負荷）制御

閉ループコントローラは、誤差 $\varepsilon(n) = Q(n) - Q_0$ の絶対値が小さくなるようにレート $E(n)$ を調整す

算出する。平衡状態において、十分なフローがソースから与えられている場合、第2段バッファへの全フローレートはリンク容量と一致しなければならない。

【0080】このように公式化していくと、 $E(n)$ の算出は制御問題となることがわかる。この問題はKolarov及びRamamurthyによる前述のABR論文に開示されているように、ABRサービスについて明示的なレート（ER）を算出する問題に、ある程度、類似している。しかしながら、スイッチ内でERを実現することは単純化できる。DRCスケジューラには、単一のコントローラで十分である。レート制御はスイッチ内で局部的に生じるため、フィードバック遅延（図6参照）はサンプリング区間 $T$ に比して小さい。この遅延は、フィードバック遅延を考慮する必要があるABRにおけるフロー制御とは異なり、無視できる。

【0081】上記した点を考慮すると、Kolarov及びRamamurthyの論文に開示された制御はノンリアルタイムABRサービスに対してのみ適用できるのに対し、本発明に係る制御は、リアルタイム及びノンリアルタイムのサービスに対して適用できることが特徴である。

【0082】コントローラの設計を単純化するために、滞留時を無限に有する単一ソースが第1段に設けられているものと仮定する（すなわち、この滞留時により、利用可能なリンク容量が常に満たされているものとする）。 $E(n)$ は時間 $n$ においてコントローラにより算出された超過レートを示すものとする。この場合、 $R(n)$ は、単一ソースに対する第1段から第2段へのフローレートである。

【0083】 $\varepsilon(n) = Q(n) - Q_0$ は、時間 $n$ におけるキューの長さ目標のキューの長さ $Q_0$ との誤差をあらわすものとする。離散的な時間に関するPD (Proportional Differential) コントローラは、一般に、次の数2式によってあらわすことができる。

【0084】

【数2】

一タフィルタを用いることができ、この場合、(1)式は次式のように簡略化できる。

【0086】

る。しかしながら、全入力トラフィック $R(n)$ における変動は閉ループコントローラの動作速度も速いことがある。この場合、目標値 $Q_0$ に近くなる前に、キューの長さ $Q(n)$ が大きな値に増大することがある。このような状態は、最低保証レート $M_i$ よりかなり大きいレートで伝送を行なうコネクションにより引き起こされる。



Q(n)の値が大きくなると、最低レートに近いレートで伝送を行なっているコネクシオンの遅延性能に悪影響を及ぼすおそれがある。閉ループコントローラの応答時間は遅すぎてスケジューラの第2段における過負荷を回避することができないため、好ましい実施例では、別の過負荷制御手法が用いられる。

【0087】第2段バッファが特定のシェーピングしきい値を越えると、フィードバックシェーピング信号がDRCスケジューラに送られる。このシェーピング信号は、スケジューラに、全てのキューを最低保証レートM<sub>i</sub>でシェーピングさせ、未使用帯域幅の分配を停止させる。この動作により、輻輳を緩和できる高速過負荷制御を実現できる。従来技術のバックプレッシャ信号と異なり、本発明に係わるシェーピング信号は、輻輳が存在する場合に、最低保証レートでキューを伝送させる重要な機能を持っている。ここでも、全てのキューに第2段バッファのキューに対して全てのセル送出を停止させることを指示するストップバックプレッシャ信号が更に使用されてもよい。

【0088】特に、従来技術において、単純な二元的（停／進）バックプレッシャ制御を行なうことにより、全てのキューに対するスループットを等化できる。すなわち、各キューはC/Nスループットを達成する。さらに、二元的信号がしばしば発生すると、セル遅延変動(CDV)が大きくなり、これによって、リアルタイムコネクシオンのQoSが悪影響を受けるおそれがある。

【0089】前述した点を考慮して、好ましい実施例では、2つの信号、すなわち、シェーピング信号及びストップ信号が用いられる。シェーピング信号しきい値は、ストップ信号しきい値より低く設定され、輻輳を緩和させるとともに、最低保証レートでの送信が可能とする。ストップ信号しきい値は非常に高く設定され、シェーピング信号を用いていることから、ストップ信号はリリーフ弁のように、最後の手段として出力されるため、殆ど出力されない。

$$\sum_{i,1} M_{ij1} \leq C, \text{ for each OP } j \text{ and}$$

$$\sum_{i,1} M_{ij1} \leq C \text{ for each IM } 1.$$

最低レートにしたがって、キューフローを静的にレートシェーピングすることにより、レートを保証することができる。しかしながら、上述したように、静的スケジューリング規則のもとでは、出力ポートボトルネック部において帯域幅を利用できる場合であっても、割り当てられた最低レートより高いレートでキューを伝送することはできない。静的なスケジューリング規則では、出力ポートのボトルネック部に関連して動作が維持されない（すなわち、出力ポートでは、送信すべきキューが入力モジュールに存在する場合でも、不動作の状態になる時

#### 5. 4 大容量スイッチのためのDRCスケジューリング

##### 5. 4. 1 単一ループフィードバック

図2bの入力モジュールにおいて、セルは、トラフィッククラスおよびコアスイッチモジュールの宛先出力ポートに応じて、キューとして格納される（以下、このような構成をクラス／OPキューとして示す）。ここで、

(i, j, 1)は、クラスi及び入力モジュール1内の宛先出力ポートjに対応するキューを示すものとする

（以下では、入力モジュール内の特定のキューに言及するために、入力モジュールに番号を付けずに、単に、

(i, j)と表記する場合もある）。各クラス／OPキューは、フローレートを制御可能な仮想ソースとして取り扱われる。

【0090】コネクシオン受付制御(CAC)アルゴリズムにしたがって、各キューに対して最低保証サービスレートを与えるスイッチを構成できる。この最低レートの値は、クラスのQoS目標と、受付けたコネクシオン数と、これらのトラフィック特性により決定される。M<sub>ij1</sub>は、キュー(i, j, 1)の最低保証サービスレートを示すものとする。入力モジュールスケジューラでは、各入力モジュールキューが最低保証レート以上のスループットを達成できることを保証しなければならない。この場合、リアルタイムトラフィックに対しては、比較的高速におけるスループットが保証されなければならない。また、ノンリアルタイムトラフィックについても最低保証スループットを確保しなければならないが、この場合には、より低速におけるスループットが保証されなければならない。

【0091】最低レートの合計は、各出力ポート及び各入力モジュールにおける回線容量を越えてはならない。すなわち、次の数3式の関係を満足しなければならない。

【0092】

【数3】

間がある）から、キューフローが統計的に多重化されない。この場合、キューフロー内のコネクシオン間で統計的多重化が行なわれるだけである。

【0093】キューフロー間で統計的多重化を達成するために、本発明では、動的レート制御(DRC)スケジューリングが使用される。ボトルネック部となる可能性のある出力ポートjにおいて、超過レートE<sub>j</sub>は、出力ポートにおけるトラフィック利用率及び宛先OP<sub>j</sub>に対応する入力モジュールの全てのキューにおけるキュー長さをあらわす情報に基づいて算出される。E<sub>j</sub>を算出す



るための方法については、後述する。動的レート  $R_{ij1}$  は、次式にしたがって、キュー  $(i, j, 1)$  に割り当てられる。

$$[0094] \quad R_{ij1} = M_{ij1} + w_i N_{ij1} E_j$$

ここで、 $w_i$  は、クラス  $i$  に対して予め割り当てられた重み係数であり、 $N_{ij1}$  は、キュー  $(i, j, 1)$  に関連するアクティブコネクションの数をあらわす。リアルタイムコネクションについて言えば、 $N_{ij1}$  は、単純にキュー  $(i, j, 1)$  に割り当てられたコネクションの数である。ノンリアルタイムコネクションについては、 $N_{ij1}$  は、いくつかの実際のコネクションが不動作状態にあるため、キュー  $(i, j, 1)$  に割り当てられたアクティブコネクション数の推定値として算出される。 $w_i$  の値は、マルチクラスCACに用いられる手法に基づいて決定される。

[0095] 上記したことからも明らかな通り、レート  $R_{ij1}$  は、CACにより決定される静的部分  $M_{ij1}$  に対して、動的部分  $w_i N_{ij1} E_j$  を加えたものである。静的部分により、最低保証レートが与えられ、一方、動的部分により、キューフローが、出力ポートのボトルネック部において、公平に（重み  $w_i$  により決定される）且つボトルネック部にオーバーロードすることなく、未使用帯域幅を有効に利用することが可能になる。

[0096] DRCスケジューリングは、また、出力モジュールキュースケジューラに対しても同様に適用される。出力モジュールにおいて、セルは、トラフィッククラス及び出力回線に応じて、キューイングされる。入力モジュールスケジューリングについて前述したものと同一表記を用い、 $(i, j, 1)$  が、クラス  $i$  及び出力モジュールOM1における宛先出力回線  $j$  に対応するキューを示すものとする。OM1におけるキュー  $(i, j, 1)$  に割り当てられた動的レート  $R_{ij1}$  (チルダ付き) は、次の数4式によって決定される。

[0097]

【数4】

$$\tilde{R}_{ij1} = \tilde{M}_{ij1} + w_i \tilde{N}_{ij1} \tilde{E}_{j1}$$

ここで、 $M_{ij1}$  (チルダ付き) は、キュー  $(i, j, 1)$  についての最低保証レート、 $N_{ij1}$  (チルダ付き) は、キュー  $(i, j, 1)$  に割り当てられたアクティブコネクションの数、 $E_{j1}$  (チルダ付き) は、OM1における回線  $j$  についてのDRCレートである。

#### 5. 4. 2 デュアルループフィードバック

前述したセクションにおいて、DRCレート  $E_j$  は、出力ポートOP  $j$  におけるボトルネック部について算出され、当該DRCレート  $E_j$  がレート  $R_{ij1}$  を算出するのに用いられた。第2のボトルネック部として、レートの不一致により、出力モジュールの出力回線にボトルネックが生じる場合が考えられる。出力回線がオーバーロード状態になると、出力モジュールのキューは長くな

り、セル損失が多くなると共に過少利用率の状態になることがある。

[0098] このような問題は、図2cに示すように、出力モジュール内の出力回線に応じて入力モジュールにおけるセルをキューイングすることにより、緩和することができる。

[0099] より具体的に言えば、 $(i, j, k, 1)$  の表記は、クラス  $i$ 、宛先出力モジュール  $j$ 、宛先出力回線  $k$  (出力モジュール  $j$  内)、及び入力モジュール1のキューを示すものとする。また、表記  $(i, j, k)$  は、宛先OM  $j$  及び出力回線  $k$  におけるクラス  $i$  のキューをあらわすものとする。この場合、入力モジュール内のキューの数は係数  $L$  と共に増加する。ここで、 $L$  は出力モジュール毎の出力回線の数である。入力モジュールから出力回線のボトルネック部に第2のフィードバックループを追加して、DRCスケジューリングを行なうことにより、スイッチ性能をさらに向上させることができる。以下、この点を簡単に説明する。

[0100] 第2のフィードバックループにおいて、レート  $E_{jk}$  はOM  $j$  内の出力回線  $k$  におけるキュー内のセルの数に基づいて算出される。ここで、 $E_{jk}$  は出力モジュール  $j$  内の出力回線  $k$  に対応するボトルネック部において利用可能な帯域幅をあらわし、この帯域幅をレートとして、全ての入力モジュールIMに伝達される。これを用いて、各入力モジュールではキュー  $(i, j, k, 1)$  に対する動的レート  $R_{ijk1}$  を次式により算出する。

[0101]

$$R_{ijk1} = M_{ijk1} + w_i \min \{E_j, E_{jk}\}$$

ここで、 $M_{ijk1}$  は、キュー  $(i, j, k, 1)$  についての最低保証レートを示す。このようにして動的レートを算出することにより、出力モジュール  $j$  及び出力回線  $k$  (出力モジュール  $j$  内) における双方のボトルネック部のレートを動的に調整できる。実際には、入力モジュールIMのキューを制御する動的レートは、前述した2つのフィードバックループにおいて算出される。これら2つのフィードバックループのうち、第1のフィードバックループは任意の出力モジュールに接続されており、第2のフィードバックループは出力モジュール内の任意の出力回線に接続されている。この構成では、出力モジュールにおけるキューを短くし、且つ、高い利用率を維持することができる。すなわち、DRCスケジューリングによる制御を行なう場合、セルをキューイングする動作は大部分入力モジュールで行なわれることになる。

[0102] 6. スケジューラの設計

#### 6. 1 レートベースのスケジューリング

各セル時間内に、入力モジュールIMのスケジューラは、セルを宛先OPに送出できる次のキューを決定する。DRCスケジューリングは、各キューに対して動的

に算出されたレートに基づいて行なわれる。この動的レートを与えられると、入力モジュールIMは、処理すべき次のキューを決定する2段階アルゴリズムを実行する。

【0103】● 第1段： 仮想レートシェーピング

● 第2段： サービススケジューリング仮想レートシェーピング及びサービススケジューリングについては、後述する。

【0104】表記  $(i, j)$  であらわされるキューについて考察する。この場合、当該キューは、クラス  $i$  ( $i = 0, \dots, 7$ ) 及び宛先  $OPj$  ( $j = 0, \dots, 15$ ) によってあらわされる。当該キュー  $(i, j)$  は、次に示すパラメータを有している。

【0105】●  $TSij$ ： タイムスタンプ。これは、12ビットの整数部及び8ビットの小数部を持つ20ビットレジスタに記憶される。タイムスタンプは、キューがサービスに対してスケジューリングあるいは再スケジューリングされるたびに、更新される。 $TSij$  はゼロに初期化される。

【0106】●  $AQij$ ： 実際のキューサイズ。これは、キュー内  $(i, j)$  に含まれるセルの数である。 $AQij$  は、セルがキュー  $(i, j)$  に到着するたびに増加し、あるセル時間に処理すべきキュー  $i$  をスケジューラが選択するたびに減少する。 $AQij$  は、16ビットレジスタに記憶され、ゼロに初期化される。

【0107】●  $VQij$ ： 仮想キューサイズ。これは、キュー  $(i, j)$  に対する仮想キューに含まれるセルの数である。 $VQij$  は、8ビットレジスタに記憶され、ゼロに初期化される。好ましい実施例において、 $VQij$  は、255を越えて増加することはない。

【0108】●  $WFi j$ ： ラップアラウンドフラグ。これは、現在クロック(時刻)のサイクルを示す2ビットフラグであり、1に初期化される。

【0109】●  $Mij$ ： 最低保証レート。この量はCACにより与えられ、間隔  $IMij$  として記憶される。

【0110】●  $Jij$ ： 最低保証レートについての間隔。これは、キュー  $(i, j)$  についての最少保証レート  $Mij$  の逆元であり、12ビットの整数部及び8ビットの小数部を持つメモリに20ビットの数として記憶される。

【0111】●  $Ej$ ： DRCレート。DRCレートは  $OPj$  のグローバル仮想キューに基づき  $OPj$  において算出される。この値は、IMに記憶されない。

【0112】●  $wi$ ： クラス  $i$  についての重み係数。これは、シェーピングレート  $Rij$  の算出に用いられる8ビットの整数である。

【0113】●  $Rij$ ： 算出されたシェーピングレート。これは、キュー  $(i, j)$  のトラフィックをシェーピングするためのDRCアルゴリズムによって算出さ

れたレートである。この値は、間隔  $Iij$  として記憶される。

【0114】●  $Iij$ ： 算出されたレートの間隔。これは、キュー  $(i, j)$  についてDRCアルゴリズムにより算出された、レート  $Rij$  の逆元である。

【0115】●  $Pij$ ： スケジューリング優先権。これは、1ビットフラグである。 $Pij = 1$  は、キュー  $(i, j)$  が、仮想キューにスケジューリングする次のセルを見つけることの優先権を有することを示す。

【0116】●  $PVij$ ： 仮想キューからのサービス優先権。これは1ビットフラグである。 $PVij = 1$  は、キュー  $(i, j)$  が、仮想キューからサービスする(OPに送出する)次のセルを見つけることの優先権を有することをあらわしている。

【0117】●  $Sij$ ： シェーピング信号。この信号を設定した場合、キュー  $(i, j)$  は、MCRレートでスケジューリングしなければならない。信号  $Sij$  は、次の条件において、1に設定される。

【0118】1. キュー  $(i, j)$  は、RTタイプであり、シェーピングRT信号は、1に設定される。

【0119】2. キュー  $(i, j)$  は、RTタイプであり、ローカルVQカウントがしきい値を越える。すなわち、 $\Sigma i, j VQij \geq ThVQ$  の関係が成り立つ。

【0120】3.  $Nij \times Wi \times Ej$  の積は0である。

【0121】ここで、NRTトラフィックは、第3の場合においてのみシェーピングされる。

【0122】●  $Nij$ ： アクティブVCの数。RTキューについて、この数は単純に、CACにより登録されたVCの数である。これは、VCを全てアクティブと仮定するからである。長期にわたって休止状態となり得るNRTキュー(すなわち、UBRやABR)について、アクティブなNRTキューの数は、サブセクション5.5で説明する計数メカニズムにより推定される。

【0123】上述した好ましい実施例において、タイムスタンプはキュー毎に割り当てられるが、セル毎にタイムスタンプを割り当てることによりレートシェーピングを実行することもできる。このような実施例において、現在時刻CT以下のタイムスタンプを有する全てのセルはサービスの資格があるものとする。

現在時刻及びラップアラウンド

各入力モジュールIMは、現在時刻を記憶する12ビットカウンタCTを有している。ここで、各時間の単位は、2.4Gbpsにおける1セル時間、すなわち、175nsである。1サイクルは、212セル時間、すなわち、CT=0から始めてCTが循環する[wrap around]までに必要とする時間として定義される。CTが一巡する度毎に、各キュー  $(i, j)$  に対する  $WFi j$  フラグは、 $WFi j$  が3未満の場合、1ずつ増加する。この動作は、一セル時間内で行なうことができる。タイムスタ

ンプTS<sub>ij</sub>及びフラグW<sub>ij</sub>は、合せて、現在時刻に関してキュー(i, j)に対して保持される時間の値を示す。WFのとり得る4つの値の意味が、表2に示されている。

【0124】

【表2】

: WF <sub>ij</sub> の値	
WF <sub>ij</sub>	意味: TS <sub>ij</sub> は…
0	CTの1サイクル前
1	CTと同一サイクル内
2	CTの1サイクル後
3	CTの少なくとも2サイクル後

TS<sub>ij</sub>の値は、WF<sub>ij</sub>とともに、ラップアラウンドが生じた時でさえも、キュータイムスタンプと現在時刻の相対値を決定することを可能とする。

## 6.2 仮想レートシェーピング

仮想レートシェーピングは、各キュー(i, j)に割り当てられたタイムスタンプTS<sub>ij</sub>に基づいて行なわれる。タイムスタンプTS<sub>ij</sub>は、キュー(i, j)のフローレートがR<sub>ij</sub>かM<sub>ij</sub>に制限されるように、更新される。動的レートR<sub>ij</sub>は、輻輳が生じていない状態では、タイムスタンプ計算に用いられる。輻輳が生じている状態では、最低レートM<sub>ij</sub>が用いられる。最低レートをキュー(i, j)に対するタイムスタンプ計算に用いられた場合には、優先ビットP<sub>ij</sub>が1に設定される。

【0125】現在時刻以下のタイムスタンプを有するキューは、有効と判断される。現在時刻は、フリーランニングクロックにより指示されており、クロック間の間隔は1セル時間に等しい。各セル時間の間、入力モジュールIMスケジューラは、1に設定された優先ビットP<sub>ij</sub>を有するキュー(i, j)に与えられる優先権を備えた次の有効キューを選択する。有効キュー、すなわち、キュー(i, j)が選択されると、仮想キューカウンタVQ<sub>ij</sub>は1だけ増加する。仮想キューカウンタは、キューとして有効であるが、まだ送出されていないセルの数を計数する。AQ<sub>ij</sub>はキュー(i, j)内のセルの数を示す。仮想キュー内にスケジュールされていないアクティブなキューにまだセルがある場合、すなわち、AQ<sub>ij</sub>>VQ<sub>ij</sub>であれば、タイムスタンプTS<sub>ij</sub>は、実際のキュー内の次のセルをいつスケジューリングするかを指示するように更新される。タイムスタンプはまた、キュー(i, j)に、セルが到来すると更新される。

【0126】スケジューリングアルゴリズムにしたがって、次の2つの条件により、優先ビットP<sub>ij</sub>が設定される。

【0127】1. タイムスタンプTS<sub>ij</sub>の現在時刻からの遅れが1/M<sub>ij</sub>よりも多い。

【0128】2. OP<sub>j</sub>における輻輳を制御するため

に、最低保証レートM<sub>ij</sub>でキューフローをシェーピングする必要がある。第1のケースは、ローカルIMスケジューラがトラフィック及び算出されたシェーピングレートを維持することができないために、発生する。レート計算が適切なレートに収束するのに時間がかかるため、IMスケジューラにおけるシェーピングレートの瞬時的合計が回線レートCを超過してしまう場合がある。これは、次の数5式であらわされる。

【0129】

10 【数5】

$$\sum_{ij} R_{ij} > C$$

この状態が発生すると、複数のキューが同じセル時間内に有効となる。

【0130】第2のケースにおいて、キューは、シェーピングモードでスケジューリングされる。すなわち、キューサービスレートがR<sub>ij</sub>からM<sub>ij</sub>に変更される。シェーピングモードは、以下の条件のもとで、呼出される。

20 【0131】1. キューはリアルタイムであり、OPからのシェーピング信号あるいはストップ信号が有効である。

【0132】2. キューがリアルタイムで、宛先OPについての仮想キューカウンタVQのカウント値の部分合計値がしきい値を越える。

【0133】3. 算出された動的レートRは、最低レートMに等しい。

30 【0134】最初の2つの条件は、出力ポートにおけるリアルタイムトラフィックについて輻輳の発生を示すものである。この状態ではシェーピングモードに迅速に切り換えることにより、このような輻輳を緩和するとともに、最低レート保証を維持できる。シェーピングモードで動作するキューに優先権を与えることにより、最低レートを保証できる。有効なキュー(i, j)が第1段スケジューラにより選択されると、その仮想キューカウンタVQ<sub>ij</sub>が1だけ増加する。優先ビットP<sub>ij</sub>が1に設定されると、第2段の優先ビットPV<sub>ij</sub>が1に設定される。

40 【0135】セルが空のキュー(i, j)に加わると、キューは有効になり、次にキューが処理される時間を指示するためにタイムスタンプTS<sub>ij</sub>が再計算される。このプロセスをスケジューリングと呼ぶ。第1段スケジューラがキュー(i, j)を選択した後、仮想キューカウンタVQ<sub>ij</sub>は1だけ増加する。アクティブなキュー内に、仮想キュー内にスケジューリングされていないセルがまだある場合、すなわち、AQ<sub>ij</sub>>VQ<sub>ij</sub>の場合には、タイムスタンプTS<sub>ij</sub>は、実際のキュー内の次のセルがスケジューリングされる時間を指示するように更新される。このプロセスを再スケジューリングと呼ぶ。スケジューリングスケジューリングについてのアル

ゴリズムをフローチャート形式で図7に示す。ステップS700において、キュー(i, j)のセルが到着すると、カウンタAQijが1だけ増加する(++の表記は、1だけ増加することを示すために用いられる)。ステップS710において、AQij-VQij=1が成立するかどうかチェックされる。成立する場合には、キューのスケジューリングが可能となり、プロセスはステップS720に移行する。成立しない場合には、処理はステップS725で終了する。

【0136】可変CCTは、タイムスタンプTSijに関連する現在時刻の値を含む14ビットの整数変数である。WFijは1に初期化される。表2を参照すると、AQij-VQij=1が成立する場合には、必然的に、WFij≥1となることがわかる。

【0137】ステップS720において、CCTは、次式により計算される。

【0138】 $CCT = CT + (WFij - 1) \ll 12$   
ここで、 $\ll$ は、2進左シフト(桁送り)演算(すなわち、212の乗算)を示している。

【0139】スケジューリングアルゴリズムにおける次のステップは、CCTと13ビットの整数TSij+I  
Mijを比較することである。ステップ730におい

て、 $CCT < TSij + Jij$  (3)

なる条件が成立しない、即ち、2偽(フォールス)であるならば、キュー(i, j)の処理が遅いと考えられる。すなわち、最低保証レート(Jij)において少なくとも1区間だけ現在時刻より遅れている。このため、処理はステップS740に進み、キュートラフィックがその最低保証レートに従っているため、優先ビットPijは1に設定される。次に、ステップS750において、キュー(i, j)は、現在時刻CTにスケジューリングされる。すなわち、TSij=CTとなる。

【0140】しかしながら、ステップS730において、式(3)が真であるならば、処理はステップS760に進んで、Sijの値をチェックする。Sijの値に応じて、最低保証レートMijあるいは算出されたレートRijでキューは、スケジューリングされる。ステップS760において、Sij=0であれば、すなわち、OPがオーバーロードされていなければ、キュー(i, j)には、ステップS770において、優先権Pij=0を付すると共に、算出されたレートRijが割り当てられる。この場合、ステップS780において、 $CCT < TSij + Iij$  (4)

の条件が偽であるならば、処理はステップS750に進み、キューは現在時刻CTにスケジューリングされる。ステップS780の条件が真であれば、タイムスタンプは、ステップS790において、次式のように更新される。

【0141】 $TSij = TSij + Iij$

ステップS730において、Sij=1であると、すなわち、OPがオーバーロードになると、キュー(i, j)には、最低保証レートシェーピングされ優先権Pij=1が与えられる(ステップS775)。この場合、ステップS795において、タイムスタンプは次式にしたがって更新される。

【0142】 $TSij = TSij + Jij$

また、ラップアラウンドフラグWFijは、ステップS755において、適切に調整される必要がある。

再スケジューリング

再スケジューリングアルゴリズムを図8に示す。このアルゴリズムは、現在時刻CTに対して適合時刻を過ぎてしまったキューを処理し、必要に応じて再スケジューリングすることを目的としている。ここで、キュー(i, j)を処理するということは、その仮想キューカウンタVQijを1だけ増加することを意味する。アルゴリズムは、Pij=1の優先権を有するキューに(i, j)のラウンド・ロビンサーチを行なう。この場合、サーチは、クラスi(i=0, ..., 7)と、さらに宛先OPj(j=0, ..., 15)に対して反復的に行なわれる。図8には示されていないが、このサーチは以下のようにして、実行される。まず、第1のサーチ処理では、Fij=1で且つ、優先ビットがセットされた(即ち、Pij=1)キュー(i, j)を見つける動作が行なわれる。このサーチでキューが見つからなかった場合、第2のサーチ処理が、Fij=1をもつキューを見つけるために実行される。再スケジューリングアルゴリズムは、全てのキューが検査されるか、現在セル時間の終了を示すタイムアウトが起こるまで、実行される。

【0143】Fij=1という条件は、下記が成立する場合にのみ真である。

【0144】1. AQij>VQij。これは、実際のキュー内に、仮想キュー内にスケジューリングされていない少なくとも1つのセルがあることを意味する。

【0145】2. VQij<FF(16進)。カウンタVQijは、VQij=FF(16進)であるとき、増加を停止する8ビットカウンタである。したがって、最大256個のセルを仮想キューにスケジューリングできる。最大値に達すると、再スケジューリングのためにキューはバイパスされなければならない。すなわち、仮想キューカウンタは255を越えて増加することはできない。仮想キューカウンタの限界を1に設定した場合、これは、実際には、仮想キューを使用不能にすることになる。スケジューラは、レートシェーピングを行なうが、レート計算は、仮想キューの全体のサイズに基づくものであってはならない。

【0146】3. WFij≥2あるいは(WFij=1及びTSij≥CT)。この条件が真であるならば、キュー(i, j)はその適合時間を通過したこと、すなわち、TSijは、CTに記録されている現在時刻より

早い時点であらわしている。有効なキュー (i, j) が、ラウンド・ロビンルーブサーチ中に見出だされると、次に、 $VQ_{ij}$  を1だけ増加させる処理が行なわれる (S820)。仮想キューの優先ビットは、次式にしたがって更新される (S820)。

【0147】  $PV_{ij} = \max(P_{ij}, PV_{ij})$  すなわち、 $PV_{ij}$  は、それがすでに1に設定されていたか、あるいは、 $P_{ij}$  が設定されている場合に1に設定される。

【0148】 次に、 $AQ_{ij} > VQ_{ij}$  であれば (S830)、キューを再スケジューリングする必要があり

(S840)、そうでなければ、再スケジューリングの必要はない (S815)。再スケジューリングステップにおいて、一時的な変数である CCT がスケジューリングアルゴリズムと同様に算出される (図7参照)。

【0149】  $CCT = CT + (WF_{ij} - 1) < 12$   $CCT < TS_{ij} + J_{ij}$  が偽であれば、キュー (i, j) は、現在時刻に遅れていると判定される (S850)。したがって、現在時刻に追いつくために、キューは、最低保証レート  $M_{ij}$  で、1に設定された優先ビット  $P_{ij}$  を付して、スケジューリングされる (ステップ S865 及び S875)。一方、 $CCT < TS_{ij} + J_{ij}$  が真であれば、 $S_{ij}$  の値はステップ S860 においてテストされる。 $S_{ij} = 0$  のとき、キューは、 $P_{ij} = 0$  を付してレート  $R_{ij}$  でスケジューリングされる (ステップ S870 及び S880)。そうでない場合、処理は、ステップ S865 及び S875 に進み、キューは、優先権  $P_{ij} = 1$  が付され、最低保証レートでスケジューリングされる。

【0150】  $TS_{ij}$  が  $J_{ij}$  あるいは  $I_{ij}$  を加えることにより更新されると、その結果、オーバーフロービット  $Z_{ij}$  が生じる。 $Z_{ij} = 1$  の場合、タイムスタンプ  $TS_{ij}$  は次のサイクルに進み、 $WF_{ij}$  は1だけ減少されることになる。そうでなければ、 $WF_{ij}$  はそのままの状態を維持する。この動作は、ステップ S890 で行なわれる。

### 6. 3 サービススケジューリング

各セル時間中には、高々1個のセルが、入力モジュールからコアTDMバス上を宛先出力ポートに送出される。図9に示されるように、セルを送り出すべきキューは、優先ビット  $PV_{ij}$  に基づいて、ラウンド・ロビンサーチにより、決定される (ステップ S900)。第2段スケジューラにおいて、キューは、 $VQ_{ij} > 0$  で、且つ、キューについての宛先OPバッファがストップモードにない場合、サービスを受け得るものと判定される。サービスを受け得るキューが見つかった (ステップ S900 で yes) と、キュー内の最初のセルがTDMバス上を伝送される (ステップ S910)。また、ステップ S910 において、 $VQ_{ij}$  及び  $AQ_{ij}$  の両方が1だけ減少し、仮想キュー優先ビット  $PV_{ij}$  がゼロにリセ

ットされる。再び、 $VQ_{ij}$  の値は、サービスを受け得るセルの数を示していることは注意すべきである。

### 6. 4 ハードウェア構成

図10は、スケジューリング動作を実行するためのハードウェアの構成を示している。主要な構成要素は、以下のとおりである。

【0151】 1. タイムスタンプ  $TS_{ij}$  用の記憶装置。これらは、複数個のレジスタ100として実現することができる。

【0152】 2. コンパレータ110のアレイ。キュー (i, j) に関連づけられたコンパレータは、 $TS_{ij}$  と  $CT$  を比較する。

【0153】 3. 実際のキュー  $AQ_{ij}$  用の記憶装置。これはカウンタ200のアレイとして実現できる。

【0154】 4. 仮想キュー  $VQ_{ij}$  130用の記憶装置。

【0155】 5. コンパレータの出力に対して優先権をラウンド・ロビン (PRR) 式にサーチするブロック135 (仮想レートシェーパー)。

【0156】 6. 仮想キューについて PRR サーチを実行するブロック145 (サービススケジューラ)。

【0157】 7. 計算エンジン150。

【0158】 8. コアスイッチからの「stop/shape/go」信号。

【0159】 図示された仮想レートシェーパー135は、キュー (i, j) に対する優先ビット  $P_{ij}$  を用いて、仮想レートシェーピングを行なう。仮想レートシェーピングの際、仮想レートシェーパー135は、 $P_{ij} = 1$  の優先権をもち、 $TS_{ij} \leq CT$  が成立するキュー (i, j) を探す。 $AQ_{ij} > VQ_{ij}$  であれば、仮想キュー  $VQ_{ij}$  は、1だけ増加する。

【0160】 サービススケジューリングのための PRR サーチを行なうサービススケジューラ145は、 $PV_{ij} = 1$  の優先権を持ち、 $VQ_{ij} > 0$  の仮想キューをラウンド・ロビン式にサーチする。仮想キュー (i, j) は、宛先出力ポート j に対応するストップ信号 (stop) が不在の場合にのみ、有効となる。

【0161】 計算エンジン150は、DRC スケジューリングに応じて、レート  $R_{ij}$  を動的に更新する。レート  $R_{ij}$  は、次式にしたがって算定される。

【0162】  $R_{ij} = M_{ij} + w_i N_{ij} E_j$

ここで、式中の値は次のとおりである。

【0163】 ● CAC からの情報：

— 最低保証レート  $M_{ij}$

— クラス重み  $w_i$

● コアスイッチモジュールからのストップ/ゴー/シェーブフィードバック。

【0164】 ● キュー (i, j) に関連づけられた、アクティブなコネクションの推定数  $N_{ij}$ 。

【0165】 ● 出力モジュール j からの I/RM セル中

に含まれる超過レート $E_j$ 。計算エンジンは、さらに、サブセクション6.2で説明したスケジューリング及び再スケジューリングアルゴリズムにしたがって、タイムスタンプ $TS_{ij}$ を更新する。

#### 6.5 アクティブVCの推定数

キュー $(i, j)$ に対するアクティブVCの数 $N_{ij}$ は、レート $R_{ij}$ の計算と、出力モジュールで算出されるER値に用いられる。リアルタイムコネクションについては、アクティブVCの数は、CACアルゴリズムにより受け付けたVCの数であると見なされる。UBRやABRなどのノンリアルタイムコネクションについては、CACによって受け付けたVCの数は、任意の時間におけるアクティブなVCの実際の数よりはるかに大きい場合がある。これは、一般に、ノンリアルタイムVCではQoSを保証する必要がなく、長期間にわたって空き状態にある場合もあるからである。

【0166】そこで、ノンリアルタイムトラフィックについてVC数を推定する方法が必要とされる。好ましい

$$\bar{N}_{ij}(n) = \epsilon N_{ij}(n) + (1 - \epsilon) N_{ij}(n-1)$$

ここで、 $\epsilon \in (0, 1)$ 。

#### 7. レート計算

##### 7.1 DRCレート

##### 7.1.1 単一フィードバックループ

単一フィードバックループの一般的構造を図2bに示す。それぞれ入力モジュールIM及び出力モジュールOMにおいてDRCスケジューリングを行なうためのレート値 $E_j$ （出力モジュールjに対応）と $E_j$ （チルダ付き）（出力回線jに対応）は、0.5msごとに1回計算される（セクション5参照）。

【0168】ここでは、DRCレートEの算出方法について説明するが、E（チルダ付き）も同様にして算出される。図11は、DRCスケジューリングレートの計算のためのフローチャート図を示す。図11において、E(n)は、n番目（0.5ms）のサンプリング区間において算出された一般的なDRCレート値をあらわす。記号VS(n)は、ボトルネック部に対応する仮想キューサイズの合計をあらわす。DRC値 $E_j$ に関連したVS(n)は、全ての入力モジュールを介して、出力ポートjに送られる全ての仮想キューの合計をあらわしている。同様に、NS(n)は、全ての入力モジュールを介して、出力モジュールjに送られるアクティブVCの総数をあらわす。

【0169】DRC値 $E_{jk}$ に関連したVS(n)は、出力モジュールjと出力回線kに対応する全ての仮想キュー合計をあらわす。この場合、NS(n)は、出力モジュールj内の出力回線kに送られるアクティブVCの数を示している。

【0170】この例では、OPボトルネック部における仮想キューの全体の長さに基づいて、Eを算出するため

実施例の40Gスイッチでは、単純なVCテーブルルックアップ法が用いられる。このテーブルには、各ノンリアルタイムVCに対して、1ビットエントリ（ゼロに初期化される）と、キュー識別子 $(i, j)$ とが保持されている。時間は、長さTsをもつ区間に分割される。VCkに属するセルがある区間において到着する場合、対応するテーブル上のエントリがゼロであると、このエントリがセット状態となり、カウント $N_{ij}$ は1だけ増加する。あるいは、テーブル上のエントリがすでにセットされている場合には、何の動作も行なわれない。当該区間の終了の際、 $N_{ij}$ は、その区間のアクティブVCの数の推定値をあらわしている。次の区間がスタートする前に、この値 $N_{ij}$ は全てクリアされる。アクティブVC数のより平滑化された推定値は、指数平均により次の数6式により求められる。

【0167】

【数6】

に閉ループ比例微分制御器（コントローラ）が用いられる。OPチャネル利用度が値U0（95%）を越える（ステップS1110参照）と、OPボトルネック部に対応する仮想キューの全長を目標値N0の近くに保つように、制御器によりEの値を調整する。OPチャネル利用度がU0より低い場合、制御器は、利用度がU0に近づくようにEを調整する。

【0171】CT(n)は、n番目の走査期間中にOPの出力において観察されるセルの数のカウント値を示す。Cが1つの走査期間中のセル時間の数であれば、n番目の期間における利用度は $U(n) = CT(n) / C$ で算出される（ステップS1100）。V(n)は、n番目の期間中のOPに対応する仮想キュー全長を全てのIMについて合計した値を示す。U(n) > U0であるならば、誤差は次式により算出される。

$$D(n) = V(n) - N0$$

ここで、N0は、目標仮想キューの全長である。さもなければ、誤り信号が目標の利用度 $C0 = U0C$ に基づいて算出され、誤り信号は次式により算出される。

$$D(n) = CT(n) - C0$$

各走査期間中に、ボトルネックレートが下記の比例微分（PD）制御に関する式を用いて算出される。これは、誤差をゼロにすることを目的としている（ステップS1140）。

$$E(n+1) = E(n) - \alpha_0 D(n) - \alpha_1 D(n-1)$$

上式において、係数 $\alpha_0$ 及び $\alpha_1$ は、システム安定性と高速の応答時間を保証するように設計された定数である。本発明者らにより実施されたシミュレーション実験

において、定数は、 $\alpha 0 = 1.25$  及び  $\alpha 1 = -0.75$  に設定された。レートがゼロより大きくなければならないという条件は、次式の演算により確保される。

【0175】

$$E(n+1) = \max \{E(n+1), 0\}$$

レート値は、また、ボトルネック回線レートによっても限定されなければならない。これはすなわち、次式であらわされる。

【0176】

$$E(n+1) = \min \{E(n+1), 0\}$$

レートは、 $[cells / 0.5ms]$  の単位で算出される。誤り信号  $D(n)$ 、 $D(n-1)$  及び値  $E(n)$  は、次のレート計算のために  $D(n-1)$  として格納される（ステップ S1150）。

#### 7. 1. 2 デュアルループフィードバック

デュアルループを用いてフィードバックを行なう場合、出力回線にしたがって、入力モジュール内にセルをキューイングすることが必要である（図2c）。また、カウンタは各出力回線に対してキューイングされたセル数を保持しなければならない。ここで、 $AQ_{jk}$  は、出力モジュール  $j$  の出力回線  $k$  に対してキューイングされたセルの数を示すものとする。

【0177】この場合、DRCレート  $E_{jk}$ （出力モジュール  $j$  内の出力回線  $k$  に対応する）は、0.5ms 毎に1回算出される（前述したセクション5.4.2参照）。 $E_{jk}$  の計算は、単一ループの場合に説明した  $E_j$  の計算と同様である。しかしながら、この場合、実際のキューサイズ  $AQ_{jk}$  は、図11において  $VS$  で示した仮想キューカウント値のかわりに用いられる。キューサイズ  $AQ_{jk}$  は、次に説明するように、出力モジュール  $j$  内の出力回線  $k$  に対するABR明示レートの計算にも用いられる。

#### 7. 2 ABR明示レート7. 2. 1 出力モジュールボトルネック

ABRサービスクラスについて、明示レート（ER）値は実際のABRクラスキューのサイズに基づいて算出される。ここで説明するABRレート計算の方法は、上記に引用したKolarovとRamamurthyによるABRサービスの論文において展開された方法にある程度類似している。但し、より高速の2.4Gbps回線速度を扱うための修正とスイッチ方式の実現のための修正が加えられている。さらに、ABRレート計算は、0.5ms 毎に1回実行される。各宛先OPについて、ER値  $ER_j$  ( $j = 1, \dots, 16$ ) が算出される。

【0178】図12には、明示レートERの計算に関するフローチャートが示されている。このフローチャートは、出力モジュールOMボトルネック及び出力回線ボトルネックの両方に適用される。 $Cabr(n)$  は、 $n$  番目の0.5ms 期間中に到着するABRセルの数を示す。ステップS1200において、 $n$  番目の期間中のA

BRに対する利用度は次式のとおりに算出される。

$$[0179] Uabr(n) = Cabr(n) / C$$

ここで、 $C$  は、ボトルネックレートにおける0.5ms 期間内のセル時間の総数である。

【0180】 $AS(n)$  は、 $n$  番目の期間についてのボトルネック部（出力モジュールあるいは出力回線）に対応する実際のABRキューのサイズを示す。すなわち、 $AS(n)$  は、あるボトルネック部に向けられた全てのABRキューについての実際のキューサイズの合計である。期間  $n-1$  における値  $AS(n-1)$  がメモリに記憶されている。この状態で、 $AS(n)$  と  $AS(n-1)$  の差がしきい値（ステップS1210において150個のセルとして例示されている）を越えたものとする。このことは、ABRキューが急速に大きくなり、高速制御を用いなければならないことを示している。したがって、ステップS1215において、IRRフィルタがアクセスされる。IRRフィルタは、また、 $AS(n)$  がしきい値  $Thigh$  を越えたとき（ステップS1220）、あるいは、フラグ  $F=1$  である（ステップS1230）ときにも、アクセスされる。

【0181】ステップS1240において、ABRトラフィックの利用度が目標より少ないと判断すると、処理は、ステップS1250に進む。他方、多い場合には、ステップS1245に進み、低利得フィルタが用いられる。ステップS1250において、実際のABRセルの合計が、低い方のしきい値  $Tlow$  より少ないと判定されると、処理はステップS1255に進み、ここで高利得フィルタが用いられる。しきい値  $Tlow$  以上であれば、処理はステップS1245に戻り、低利得フィルタが用いられる。

【0182】図13は、IRRフィルタの動作を示す。IRRフィルタには、DRCローカルレート  $E$  の何分の一かに等しいERレートが単純に設定される。この場合、次式の関係が成立する（ステップS1310）。

$$[0183] ER(n+1) = E(n) / 2$$

IRRフィルタは、 $AS(n)$  の値がしきい値  $Tlow$  より大きいのか小さいかに応じてフラグ  $F$  を設定あるいはリセットする（ステップS1320）。

【0184】図12から、 $F=1$  の場合に、IRRフィルタはアクセスされることがわかる。これにより、ABRトラフィックに対しては厳しい制限が加えられることになる。図13において、誤り信号  $D(n-1)$  は、IRRフィルタにおいて使用されなくても更新され、記憶される。

【0185】図14は高利得フィルタの動作を示す。主要な制御式は次の数7式のとおりである。

【0186】

$$[数7] ER(n+1) = ER(n) - \alpha 0 D(n) / NSabr(n) - \alpha 1 D(n-1) / NSabr(n-1)$$

ここで、 $NSabr(n)$  は、あるボトルネックに対応するアクティブABRであり、VC全ての合計の推定値であり、ABRクラス重み $Wabr$ により重み付けされた値である。フィルタ係数の値は、ローカルDRCフィルタにおけるものと同じである。すなわち、 $\alpha 0 = 1$ 、 $\alpha 1 = -0.75$ である。高利得フィルタについて、フィルタ係数は、 $NSabr$ によって決定される。

【0187】フィルタにおける動作は以下のように行われる。まず、実際のキュー長さ $D(n)$ と目標長さの差が決定される。ステップ1410において、ステップ1400で計算された差から、高利得フィルタが使用される。ステップ1420において、 $D(n)$ は、次回に備えて

$$\begin{aligned} ER(n+1) = & ER(n) - \alpha 0 D(n) - \alpha 1 D(n-1) \\ & - \beta 0 ER(n) - \beta 1 ER(n-1) \\ & - \dots - \beta 10 ER(n-10) \end{aligned} \quad (6)$$

低利得フィルタの係数は $NSabr$ によって決定されない。低利得フィルタのための係数値は、表3に示されている。

【0190】

【表3】

低利得ABRフィルタについての係数値

係数	値
$\alpha 0$	0.0627
$\alpha 1$	-0.0545
$\beta 0$	0.8864
$\beta 1$	0.0955
$\beta 2$	0.0545
$\beta 3$	0.0136
$\beta 4$	-0.0273
$\beta 5$	-0.0682
$\beta 6$	-0.1091
$\beta 7$	-0.1500
$\beta 8$	-0.1909
$\beta 9$	-0.2318
$\beta 10$	-0.2727

低利得フィルタにおける処理動作は、利得式が異なることを除き、高利得フィルタの処理動作と同じである。したがって、図15に示された低利得フィルタの処理動作の説明はここでは省略する。

### 7.3 制御情報の伝送

全てのDRCレートとABR ERレートの計算はそれぞれの出力モジュールOMにおいて実行される。各走査期間において、各入力モジュールIMは、全ての出力モジュールOMにキューの長さをあらわす情報を伝送する。この情報は、内部リソース管理（IRM）セルと呼ばれる特別の制御セルによって伝送される。これらのセルは入力モジュールIMにより生成され、制御シグナリングオーバーヘッドを構成する。

【0191】キュー長さ情報に基づいて、各OM $j$ は、局所的な制御を行ない、この制御ではDRCレート $E_j$ を算出する一方、ABRソース制御の際に、明示レート（ER） $ER_j$ を算出する。ABR ER値は、ソースの方向に向かうリソース管理（RM）セルにより、遠隔

$(n-1)$ に置き換えられる。ステップ1430において、 $\max\{ER(n+1), 0\}$ の動作を行なうことによって、 $ER(n+1)$ が負ではないことが保証される。更に、 $\min\{ER(n+1), E(n)\}$ の動作を行なうことによって、 $ER(n+1)$ がローカルDRCレート $E(n)$ 以下であることを保証する。ステップ1440において、全てのER値は、時間的にシフトされる。

【0188】図15に示された低利得フィルタにおける制御式は次の数8式のとおりである。

【0189】

【数8】

ABRソースに伝送される。同様に、出力モジュールOMにより生成されたIRMセルは、入力モジュールIMにDRCレート情報を伝送するために用いられる。

【0192】8. バッファ管理

20 スケジューラと共に動作する各入力モジュールIM及び出力モジュールOMは、バッファ割当ての機能を持つキューマネージャーを備えている。好ましい実施例による大容量スイッチ構造において、出力モジュールOMバッファは、出力回線ボトルネック部における競合から生ずる輻輳を扱い、一方、入力モジュールIMバッファは、OPボトルネック部における競合から生ずる輻輳を扱う。入力モジュールIM及び出力モジュールOM内のキューマネージャーは、独立したものであるが、同様な構造を備えている。入力モジュールIM及び出力モジュールOM内のセルバッファは、全てのキュー間で共有されるが、最大キューサイズには制限がある。

【0193】各キューには、トラフィッククラス及びQoS要求に基づいて予め割当てられたセル廃棄しきい値が設定されている。廃棄しきい値は、サイズ増加順に下記のとおり挙げられる。

【0194】● CLP=1であるセルを廃棄。

【0195】● 初期パケット廃棄（EPD）。新たなパケットに属するセルを廃棄。

【0196】● 部分パケット廃棄（PPD）。全てのセルを廃棄。キューマネージャーは、最低保証レートにシェーピングされている、あらゆるキューフロー中におけるCLP=1であるセルを廃棄する。このようにして、CLP=0であるトラフィックには最低保証レートが割り当てられる。

【0197】9. 性能評価

DRCスケジューリングの主要な目標は、スイッチにおけるボトルネックレートを整合させて、輻輳回避と、高効率の維持という2つの目的を満足させることである。さらに競合するクラス間で公平に未使用帯域幅を配分することをも目標の一つとしている。このセクションで



は、スイッチ設計におけるDRCスケジューリングの主要な性能特性を強調するために、いくつかの代表的なシミュレーション結果を提示する。

#### 9.1 レート制御の収束

図16を参照して、コアスイッチモジュール上の同一の出力ポート1に送出される2つのフローが与えられた場合におけるスイッチの動作を説明する。

【0198】1. 一定の入力レート0.58及び最低保証レート $M1=0.6$ を有する、IM1上のCBRフロー。

【0199】2. 一定の入力レート0.9及び最低保証レート $M2=0.3$ を有する、IM2上のUBRフロー。この場合、UBRフローに対しては、その最低保証レートが守られないことがある。これは、UBRソースがネットワーク端において監視されないことから、発生する。対照的に、CBRソースは、その最低保証レートより低いレートで実際に伝送を行なっている。

【0200】フロー $i$ のDRCレートは、 $R_i = M_i + E$ で算出される。ここで、 $E$ は、閉ループ制御により算出される利用可能な未使用帯域幅である。時刻0において、帯域は空いているから、初期的に、 $E=1$ である。すなわち、2つのフローが時刻0において同時に生成されると、各フローは、まず回線レート、すなわち、 $R_i(0+) = 1$  ( $i=1, 2$ )で伝送される。

【0201】時刻 $t=0+$ において、OP1への総フローレートは1.48である。したがって、OP1においてバッファリングが行なわれる一方、全仮想キューが入力モジュールにおいて形成される。DRCメカニズムは、DRCレート $E$ を低下させることにより動作を開始する。

【0202】図17は、フローレート $R_i(t)$ のグラ

CBRフロー1及びUBRフロー2についての遅延性能

遅延量 [セル時間]	CBRフロー1	UBRフロー2
平均遅延	0.90 ± 0.066	2.54 ± 0.71
標準偏差遅延	0.03	9.09
平均interdeparture	1.51 ± 0.11	1.61 ± 0.11
標準interdeparture	0.67	4.28

表4において、interdepartureは、セルが出力回線から出力されるときセル間隔であり、例えば、セルAAとBの出力時刻がそれぞれ $t_1$ 及び $t_2$ であるものとする、interdeparture $f$ は、 $t_2 - t_1$ となる。

【0206】ここで、フロー1をCBRクラスからUBRクラスに変更する場合を考慮する。UBRフローとして、フロー1は、フロー2とともに、OP1のノンリアルタイムバッファにおいてバッファリングされる。この例のシミュレーション結果を表5に示す。表5における両方のフローの平均遅延量は、表4における対応する遅延量に対して増加していることがわかる。特に、CBRフローとしてのフロー1の遅延量は全て、UBRフローとしての対応する遅延量よりも著しく良好であることがわかる。この例は、スイッチ構造が、ノンリアルタイムトラフィックに比べてリアルタイムトラフィックに対して提供するよりきびしいQoS制御を行なうことを証明している。

フを示している。図から、レートが比較的迅速に(約6msで)定常状態値に収束することがわかる。CBRフロー $R1(t)$ は0.58の帯域幅を用い、UBRフロー $R2(t)$ はレート0.9で入力モジュールIM2に入力される。しかし、UBRフローには、0.3のスレープットしか保証されていない。したがって、DRCレート $E$ についての正しい値は、0.12である。このため、レートは次式のように収束する。

【0203】 $R1(t) \rightarrow 0.72$ 及び $R2(T) \rightarrow$

0.42 CBRフロー $R1(t)$ はレート0.72で出力ポートOP1に伝送できるが、入力モジュールIM1にはレート0.58で入力される。一方、UBRフロー $R2(t)$ は、入力モジュールIM2においてレート0.42にシェーピングされる。UBRセルは、セルバッファ容量を超過した後、IMにおいて廃棄される。

9.2 リアルタイム対ノンリアルタイム遅延性能遅延性能を調べるために、フロー2をランダムなオン・オフ期間を有するUBRフローに置き換えることにより、上記の例を修正する。オン・オフ期間は、それぞれ、平均8及び12 [セル時間]を用いて指数的に分散される。オン期間中、フロー2では、一定レート0.93で入力モジュールIM2にセルが与えられるものとする、フロー2の平均レートは、0.372である。

【0204】また、 $M2=0.38$ を設定する。シミュレーションから得られた遅延量を表4に示す。平均遅延量は、対応する98%信頼区間とともに、セル時間の単位で与えられる。CBRフローがごくわずかな遅延と遅延ジッタを受けることがわかる。

【0205】

30 【表4】

【0207】

【表5】

UBRフロー1及びUBRフロー2についての遅延性能

遅延量 [セル時間]	UBRフロー1	UBRフロー2
平均遅延	1.79±0.42	1.69±0.22
標準偏差遅延	5.0	3.92
平均interdeparture	1.5±0.11	1.62±0.12
標準interdeparture	1.03	4.26

## 9.3 DRC対静的優先権スケジューリング

図18は、3つのオン・オフ型フローを与えられたスイッチを示している。3つのフローの仕様を表6に示す。

各フローは、トラフィッククラス、平均オン・オフ期間、オン期間時のレート、ソース入力モジュール、宛先

出力モジュール、最低保証レート (DRCに対する) に関連づけられている。

【0208】

【表6】

3つのフローの仕様

フロー No.	クラス	平均 オン	平均 オフ	オン レート	IM	OP	M1
1	Rt-VBR	12	8	0.9	1	1	0.65
2	Nrt-VBR	8	12	0.93	1	2	0.3
3	Rt-VBR	7	13	0.93	2	1	0.33

フロー1及び3は、リアルタイムVBRフローであり、フロー2は、ノンリアルタイムVBRフローである。さらに、フロー1及び2は、IM1におけるサービスについて競合する。この例では、DRCスケジューリングとIM1における静的優先権スケジューリングを比較する。静的優先権は、リアルタイムVBRフロー1に、ノンリアルタイムVBRフロー2より高い優先権を与える。明らかに、フロー1は、この方式のもとで最高の遅延性能を達成する。しかしながら、このことは、フロー2には悪影響を与える。DRCスケジューリングでは、

両方のフローにレート保証を与えることにより調整が行なわれる。

【0209】DRC及び静的優先権スケジューリングについての遅延結果をそれぞれ表7及び表8に示す。静的優先権のもとでは、フロー1は小さい遅延しか受けない。しかしながら、フロー2の遅延は、比較的大きい。DRCのもとでは、フロー1の遅延性能は、少しだけ調整されるが、フロー2の遅延性能は著しく向上する。

【0210】

【表7】

DRCスケジューリングのもとでの遅延結果

遅延量 [セル時間]	フロー1	フロー2	フロー3
平均遅延	28.0±3.68	25.6±3.62	34.9±5.53
標準偏差遅延	31.24	31.82	63.40
平均interdeparture	1.57±0.11	2.24±0.14	2.11±0.16
標準interdeparture	1.02	2.00	3.14

【0211】

【表8】

静的優先権スケジューリングのもとでの遅延結果

遅延量 [セル時間]	フロー1	フロー2	フロー3
平均遅延	10.5±1.81	121.5±17.4	3.25±0.33
標準偏差遅延	26.48	104.3	7.72
平均interdeparture	1.56±0.11	1.92±0.14	1.57±0.11
標準interdeparture	2.09	3.68	3.94

【0212】

【発明の効果】上記説明からわかるように、本発明によるスイッチは、異なるQoSを有するセルストリームを効率的に処理するものである。さらに、本発明によるスイッチは、効率的な優先権方式を用いて、ユニキャスト及びマルチキャスト伝送を効率的に多重化する。スイッチの入力側と、コアと、出力側にバッファを用いることにより、ボトルネックを悪化させることなく、動作維持を可能とする。さらに、シェーピングフィードバック信号が、一時的な輻輳を緩和させるとともに、最低保証レ

ートを確約するために、動作維持を一時的に停止するために用いられる。

【0213】また、本発明では、入力バッファ、出力バッファ、及び、コアバッファの3重バッファ形式を採用すると共に、フィードバック方式を採用することにより、高速の大容量バッファを必要とせず、且つ、入力ポート間における制御を軽減できる。

【図面の簡単な説明】

【図1】本発明の好ましい実施例によるコアスイッチモジュールの構造を示す図である。

【図2】(a)は、本発明の好ましい実施例によるVC毎のキューを構成する入力及び出力モジュールの構造を示す図である。(b)は、本発明の好ましい実施例による出力ポートクラス毎のキューを構成する入力及び出力モジュールの構造を示す図である。(c)は、本発明の好ましい実施例による出力回線クラス毎のキューを構成する入力及び出力モジュールの構造を示す図である。

【図3】本発明の好ましい実施例による入力及び出力モジュールの構成をより詳細に示す図である。

【図4】最小レートシェーピングをとまなうスケジューラを示す図である。

【図5】本発明によるDRCスケジューリングを実現するスケジューラを示す図である。

【図6】閉ループレート制御を示す図である。

【図7】タイムスタンプに応じてセルをスケジューリングするためのアルゴリズムのフローチャートである。

【図8】タイムスタンプに応じてセルを再スケジューリングするためのアルゴリズムのフローチャートである。

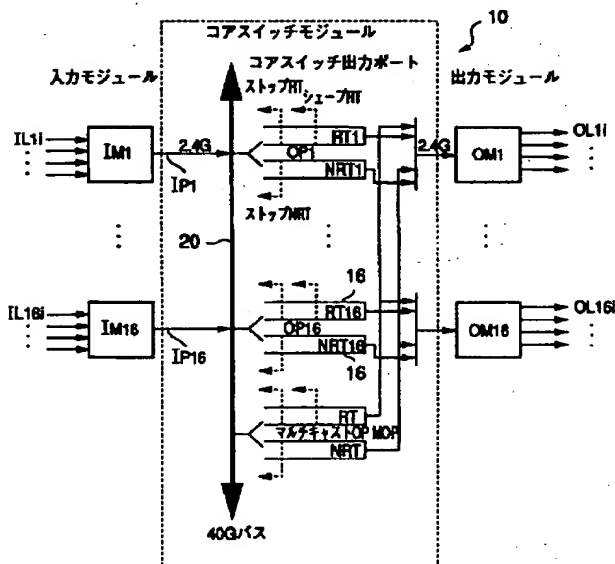
【図9】仮想キューからセルを処理するフローチャートである。

【図10】本発明の好ましい実施例によるスケジューラのブロック図である。

【図11】DRCスケジューリングについてレート計算のためのアルゴリズムのフローチャートである。

【図12】ABRについての明示的レート計算のフローチャートである。

【図1】



【図13】ABRについてのIRRフィルタのフローチャートである。

【図14】ABRについての高利得フィルタのフローチャートである。

【図15】ABRについての低利得フィルタのフローチャートである。

【図16】コアシッチの出力ポートに与えられる2つのセルストリームフローを示す図である。

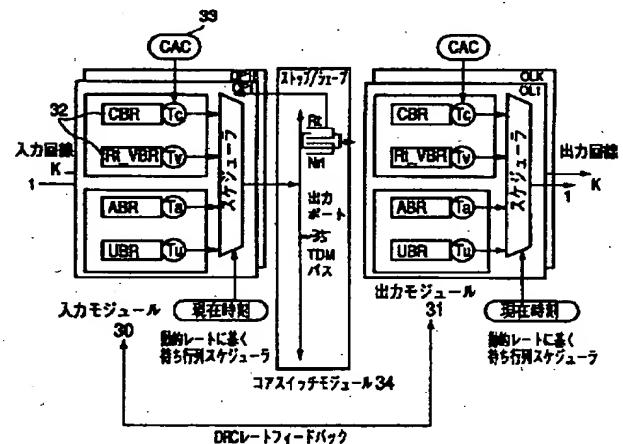
【図17】CBRフロー及びUBRフローについてDRCレートの収束を示す、図16のモデルで実施されたシミュレーションで収集されたデータのグラフである。

【図18】コアシッチの2つの出力ポートに与えられる3つのセルストリームフローを示す図である。

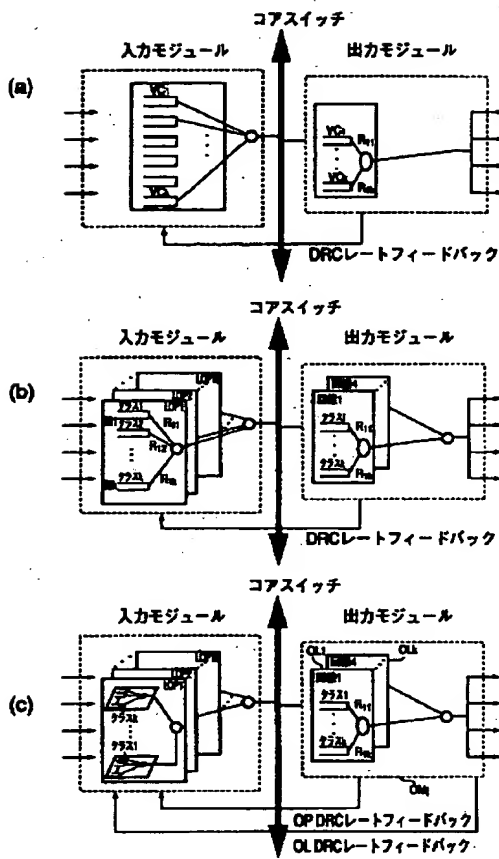
【符号の説明】

- 10、34 コアシッチモジュール
- 20、35 TDMバス
- 30 入力モジュール
- 32 バッファ
- 33 CAC
- 20 レジスタ
- 110 コンパレータ
- 200 カウンタ
- 130 仮想キューVQ<sub>i,j</sub>
- 135、145 ブロック
- 150 計算エンジン

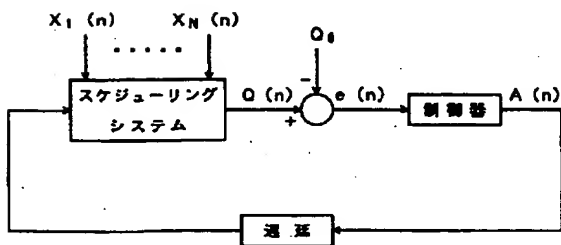
【図3】



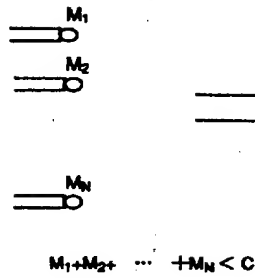
【図2】



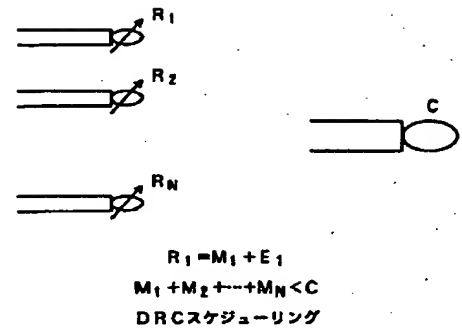
【図6】



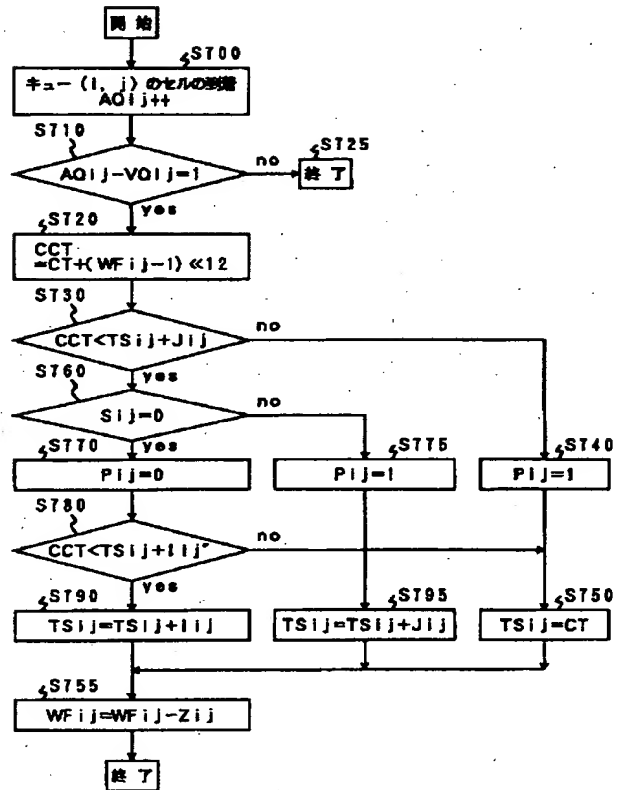
【図4】



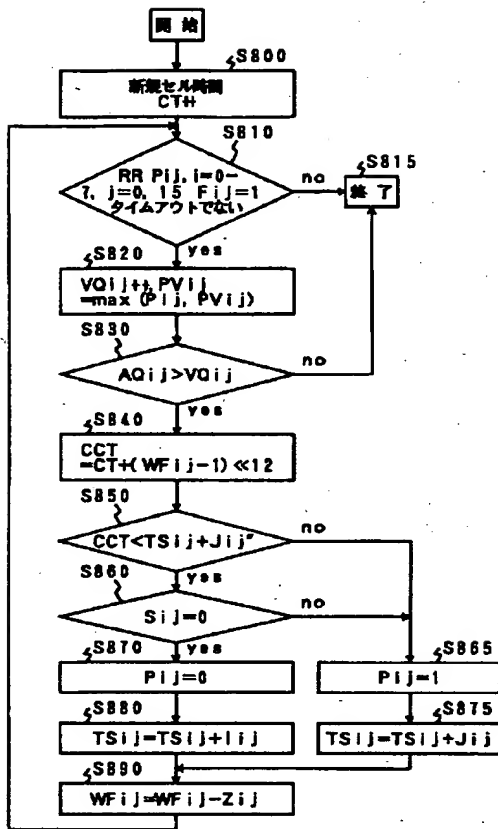
【図5】



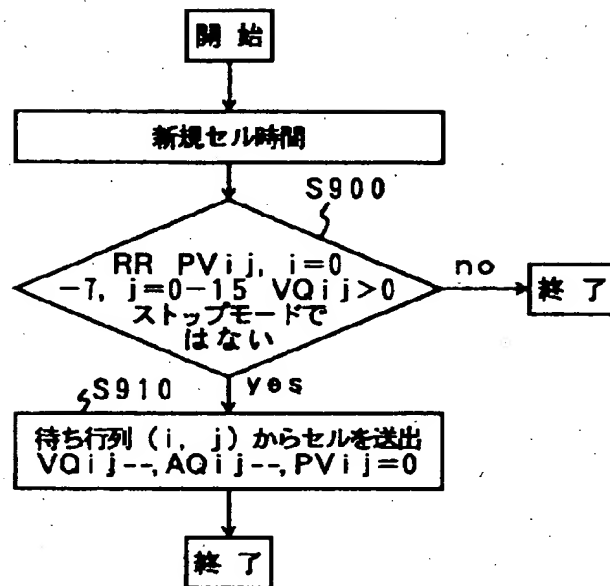
【図7】



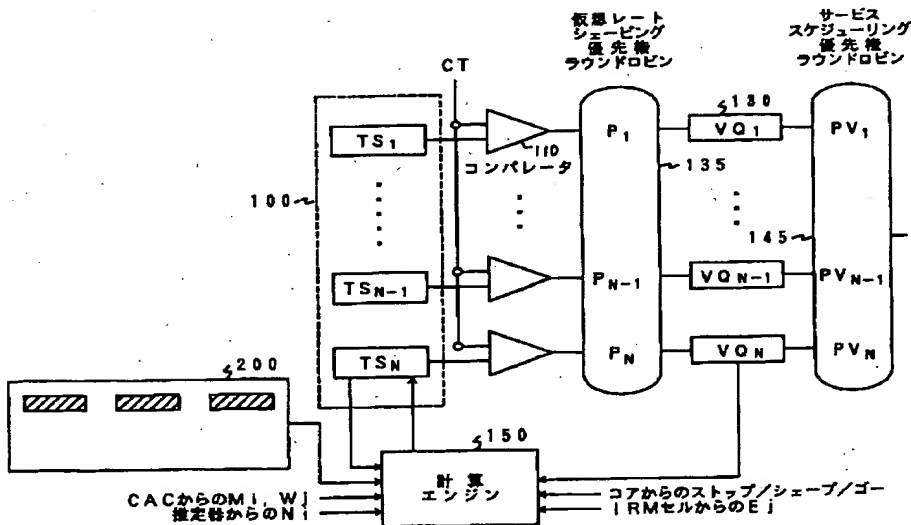
【図8】



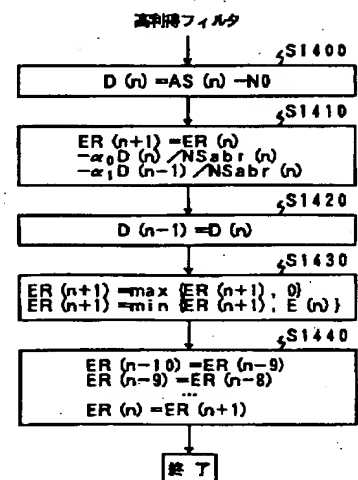
【図9】



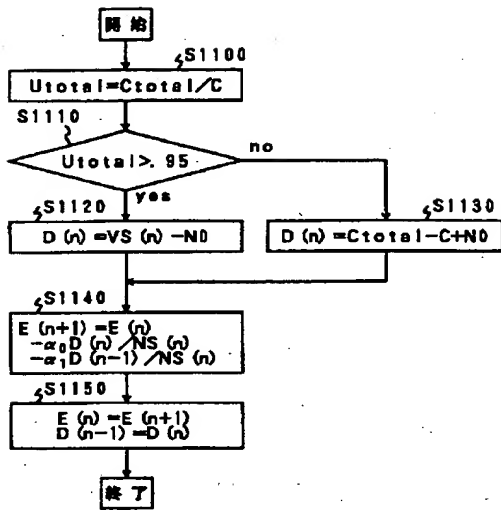
【図10】



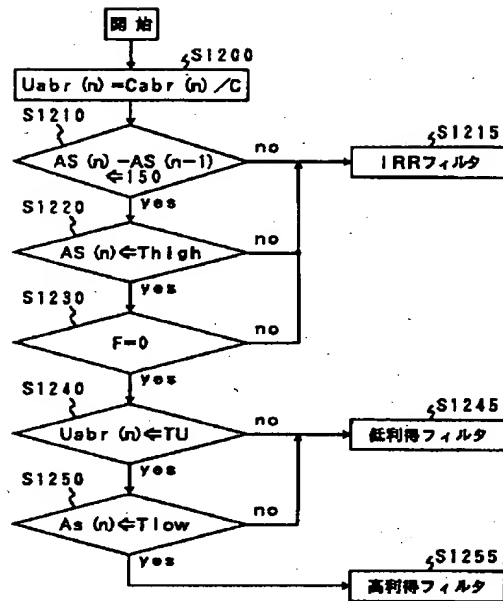
【図14】



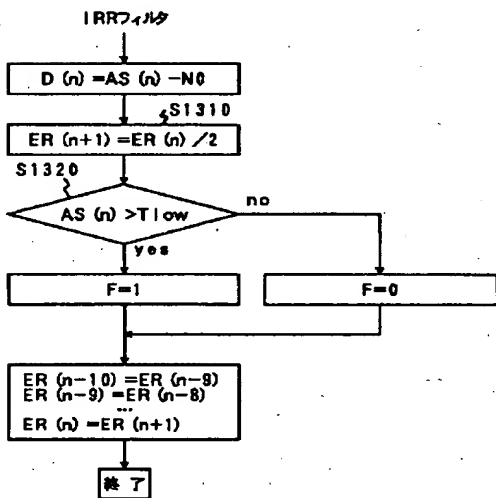
【図11】



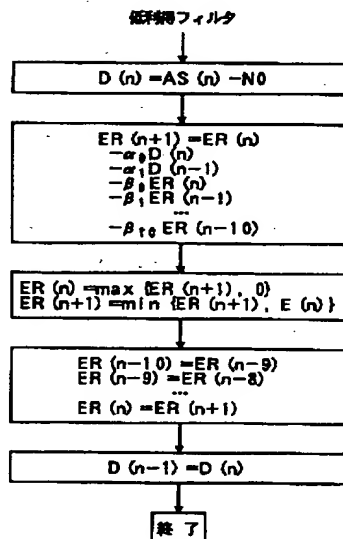
【図12】



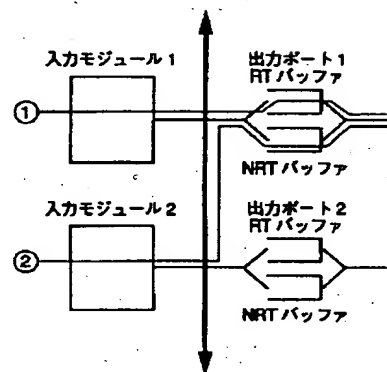
【図13】



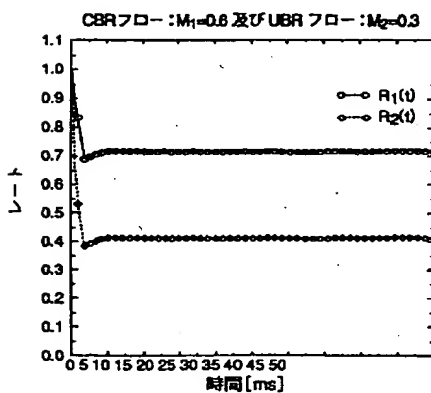
【図15】



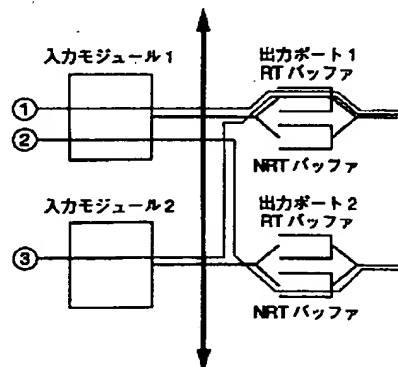
【図16】



【図17】



【図18】



## フロントページの続き

(72)発明者 ブライアン マーク  
アメリカ合衆国, ニュージャージー  
08540, プリンストン, 4 インディペン  
デンス ウエイ, エヌ・イー・シー・ユ  
ー・エス・エー・インク内

(72)発明者 ゴバラクリシナン ラママーシー  
アメリカ合衆国, ニュージャージー  
08540, プリンストン, 4 インディペン  
デンス ウエイ, エヌ・イー・シー・ユ  
ー・エス・エー・インク内